

Introduction to multilevel models

Alastair H Leyland

MRC Social and Public Health Sciences Unit
Glasgow, Scotland

Outline

- Multilevel data structures
- Algebraic formulation of multilevel models

Multilevel structures

- Strict hierarchies
 - ▲ Designs including time
 - ▲ Multiple responses
- Cross-classified structures
- Multiple membership models
- Other applications

Strict hierarchies

Hospital (2)

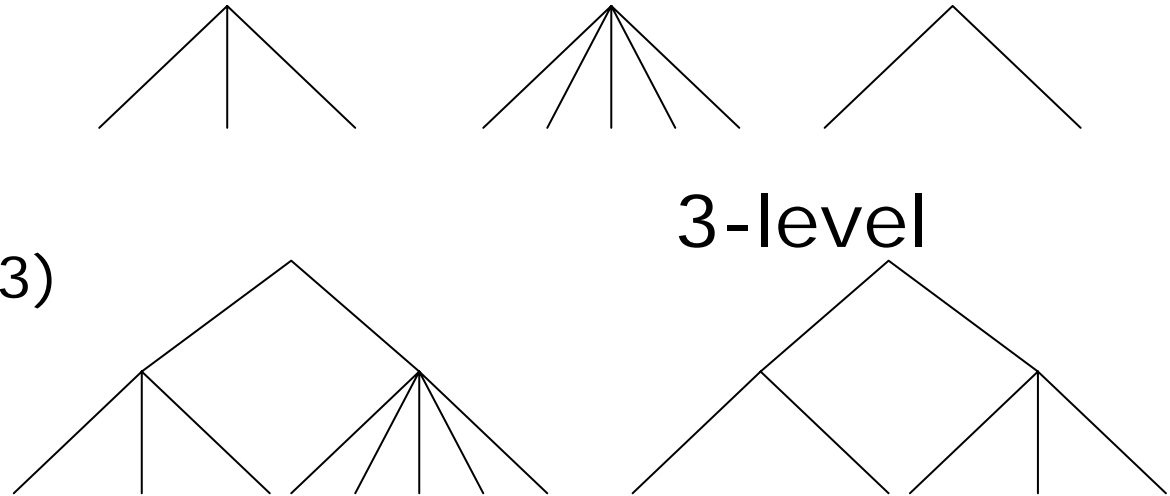
Patient (1)

Neighbourhood (3)

Household (2)

Individual (1)

2-level



...many within few

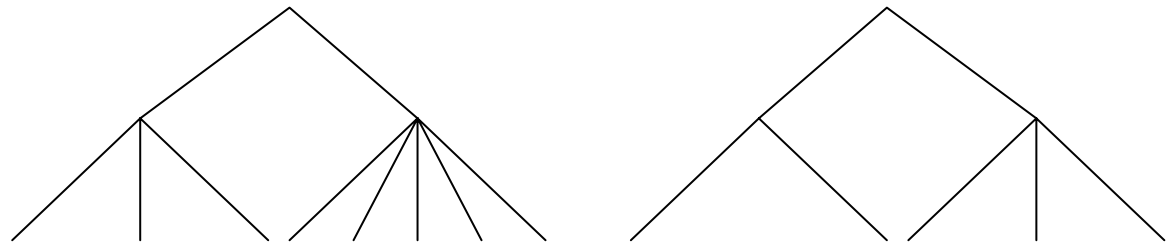
Designs including time

Repeated cross-sectional

Hospital (3)

Year (2)

Patient (1)

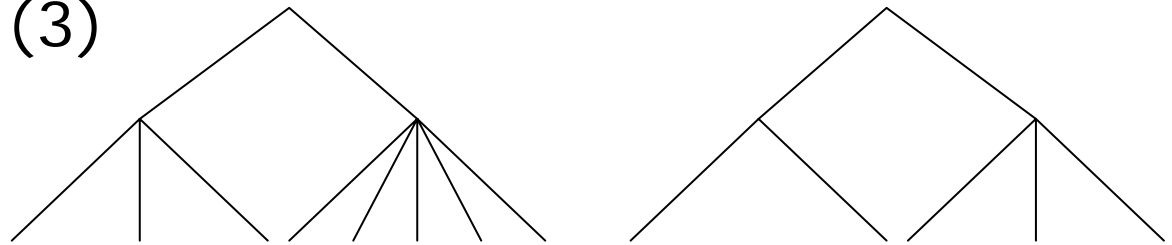


Repeated measures or panel

Neighbourhood (3)

Individual (2)

Year (1)

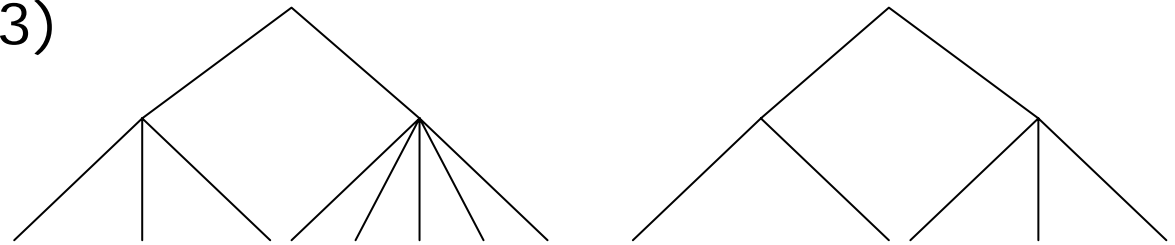


Multiple responses

Neighbourhood (3)

Individual (2)

Response (1)



Responses may be

- health-related behaviour e.g.
 - ❖ Smoking, drinking, diet, exercise
- related responses e.g.
 - ❖ Systolic and diastolic blood pressure
 - ❖ Length of stay and readmissions

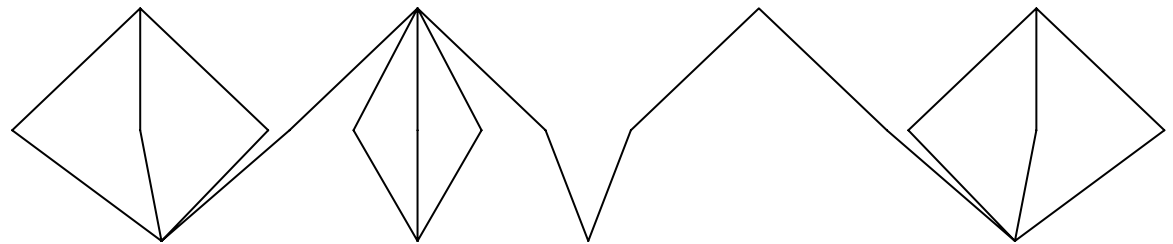
Non-hierarchical structures

Cross-classified model

Hospital (2)

Patient (1)

GP (2)



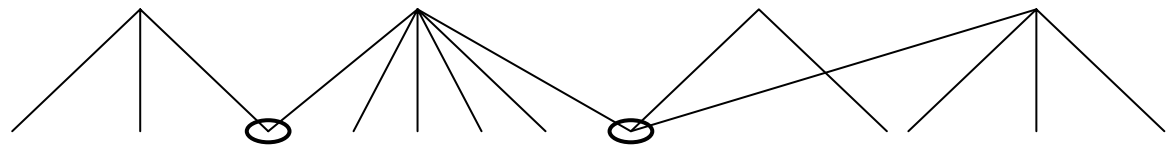
- GPs are not nested within hospitals
- Hospitals are not nested within GPs
- Patients are nested within a cross-classification of GP and hospital

Non-hierarchical structures

Multiple membership models

Hospital (2)

Patient (1)



- Some patients attend more than one hospital
- There is no strict nesting of patients within hospitals
- Can we weight the relative contribution of each hospital to the outcomes?

Algebraic formulation

OLS

$$y_i = \beta_0 + \sum_{p=1}^P \beta_p x_{pi} + e_i$$

Multilevel

$$y_{ij} = \beta_0 + \sum_{p=1}^P \beta_p x_{pij} + u_{0j} + e_{0ij}$$

$$u_{0j} \sim N(0, \sigma_{u0}^2) \quad e_{0ij} \sim N(0, \sigma_{e0}^2)$$

Residuals and variances

$$u_{0j} \sim N(0, \sigma_{u0}^2) \quad e_{0ij} \sim N(0, \sigma_{e0}^2)$$

$$\text{Cov}(e_{0ij}, e_{0i'j}) = \text{Cov}(e_{0ij}, e_{0i'j'}) = 0$$

$$\text{Cov}(u_{0j}, u_{0j'}) = 0 \quad \text{Cov}(u_{0j}, e_{0ij}) = 0$$

Intraclass
correlation
coefficient

$$\rho_I = \frac{\sigma_{u0}^2}{\sigma_{u0}^2 + \sigma_{e0}^2}$$

Variance component models

$$y_{ij} = \underbrace{\beta_0 + \sum_{p=1}^P \beta_p x_{pij}}_{\text{Fixed part}} + \underbrace{u_{0j} + e_{0ij}}_{\text{Random part}}$$

Fixed part

Random part

$$y_{ij} = \beta_{0j} + \sum_{p=1}^P \beta_p x_{pij} + e_{0ij}$$

random
intercept

$$\beta_{0j} = \beta_0 + u_{0j}$$

could be x_{pj} , $x_{pij}x_{p'ij}$

$x_{pj}x_{p'j}$ or $x_{pij}x_{p'j}$

Variances and covariances

$$y_{ij} = \beta_0 + \sum_{p=1}^P \beta_p x_{pij} + u_{0j} + e_{0ij}$$

$$\text{Var}(y_{ij}) = \sigma_{u0}^2 + \sigma_{e0}^2$$

$$\text{Cov}(y_{ij}, y_{i'j}) = \sigma_{u0}^2$$

$$\text{Cov}(y_{ij}, y_{i'j'}) = 0$$

Structure of covariance matrix

$$Var(\mathbf{Y}) = \begin{bmatrix} \sigma_{u0}^2 + \sigma_{e0}^2 & \dots & \sigma_{u0}^2 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \sigma_{u0}^2 & \dots & \sigma_{u0}^2 + \sigma_{e0}^2 & 0 & \dots & 0 \\ 0 & \dots & 0 & \sigma_{u0}^2 + \sigma_{e0}^2 & \dots & \sigma_{u0}^2 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & \sigma_{u0}^2 & \dots & \sigma_{u0}^2 + \sigma_{e0}^2 \end{bmatrix}$$

Misestimated precision

OLS
$$Var(\hat{\boldsymbol{\beta}}) = \sigma_{e0}^2 (\mathbf{X}^T \mathbf{X})^{-1}$$

ML
$$Var(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1}$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{Y}$$

Misestimated precision and ICC

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + u_{0j} + e_{0ij}$$

$$SE\left(\hat{\beta}_1\right) = SE\left(\hat{\beta}_{1,OLS}\right) \left[1 + \rho_y \rho_x \left(\bar{n}_j - 1\right)\right]^{1/2}$$

Random slopes models

$$y_{ij} = \underbrace{\beta_0 + \sum_{p=1}^P \beta_p x_{pij}}_{\text{Fixed part}} + \underbrace{u_{0j} + x_{1ij}u_{1j} + e_{0ij}}_{\text{Random part}}$$

Fixed part

Random part

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{1ij} + \sum_{p=2}^P \beta_p x_{pij} + e_{0ij}$$

$$\beta_{0j} = \beta_0 + u_{0j}$$

$$\beta_{1j} = \beta_1 + u_{1j}$$

random
slope



Random slopes models

$$y_{ij} = \underbrace{\beta_0 + \sum_{p=1}^P \beta_p x_{pij}}_{\text{Fixed part}} + \underbrace{u_{0j} + x_{1ij}u_{1j} + e_{0ij}}_{\text{Random part}}$$

Fixed part

Random part

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{u0}^2 & \sigma_{u01} \\ \sigma_{u01} & \sigma_{u1}^2 \end{bmatrix} \right)$$

Variances and covariances

$$y_{ij} = \beta_0 + \sum_{p=1}^P \beta_p x_{pij} + u_{0j} + x_{1ij} u_{1j} + e_{0ij}$$

$$Var(y_{ij}) = \sigma_{u0}^2 + 2x_{1ij}\sigma_{u01} + x_{1ij}^2\sigma_{u1}^2 + \sigma_{e0}^2$$

$$Cov(y_{ij}, y_{i'j}) = \sigma_{u0}^2 + (x_{1ij} + x_{1i'j})\sigma_{u01} + x_{1ij}x_{1i'j}\sigma_{u1}^2$$

$$Cov(y_{ij}, y_{i'j'}) = 0$$

Three level models

$$y_{ijk} = \beta_0 + \sum_{p=1}^P \beta_p x_{pijk} + v_{0k} + u_{0jk} + e_{0ijk}$$

$$Var(y_{ijk}) = \sigma_{v0}^2 + \sigma_{u0}^2 + \sigma_{e0}^2$$

$$Cov(y_{ijk}, y_{i'jk}) = \sigma_{v0}^2 + \sigma_{u0}^2$$

$$Cov(y_{ijk}, y_{i'j'k}) = \sigma_{v0}^2$$

$$Cov(y_{ijk}, y_{i'j'k'}) = 0$$

Cross-classified models

$$y_{i(j_1 j_2)} = \beta_0 + \sum_{p=1}^P \beta_p x_{pij_1 j_2} + u_{0j_1}^{(1)} + u_{0j_2}^{(2)} + e_{0i(j_1 j_2)}$$

$$Var\left(y_{i(j_1 j_2)}\right) = \sigma_{u0}^{(1)2} + \sigma_{u0}^{(2)2} + \sigma_{e0}^2$$

$$Cov\left(y_{i(j_1 j_2)}, y_{i'(j_1 j_2)}\right) = \sigma_{u0}^{(1)2} + \sigma_{u0}^{(2)2}$$

$$Cov\left(y_{i(j_1 j_2)}, y_{i'(j_1' j_2)}\right) = \sigma_{u0}^{(2)2}$$

$$Cov\left(y_{i(j_1 j_2)}, y_{i'(j_1 j_2')} \right) = \sigma_{u0}^{(1)2}$$

$$Cov\left(y_{i(j_1 j_2)}, y_{i'(j_1' j_2')} \right) = 0$$

Multiple membership models

$$y_{i\{j\}} = \beta_0 + \sum_{p=1}^P \beta_p x_{pi\{j\}} + \sum_{j=1}^{n_j} w_{ij} u_{0j} + e_{0i\{j\}}$$

$$\sum_{j=1}^{n_j} w_{ij} = 1 \quad \forall \quad i \quad w_{ij} \geq 0$$

$$Var\left(y_{i\{j\}}\right) = \sigma_{u0}^2 \sum_{j=1}^{n_j} w_{ij}^2 + \sigma_{e0}^2$$

$$Cov\left(y_{i\{j\}}, y_{i'\{j\}}\right) = \sigma_{u0}^2 \sum_{j=1}^{n_j} w_{ij} w_{i'j}$$

Multiple response models

$$y_{ij}^{(1)} = \beta_0^{(1)} + \sum_{p=1}^P \beta_p^{(1)} x_{pij} + u_{0j}^{(1)} + e_{0ij}^{(1)}$$

$$y_{ij}^{(2)} = \beta_0^{(2)} + \sum_{p=1}^P \beta_p^{(2)} x_{pij} + u_{0j}^{(2)} + e_{0ij}^{(2)}$$

$$\begin{bmatrix} u_{0j}^{(1)} \\ u_{0j}^{(2)} \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{u0}^{(1)2} & \sigma_{u0,u0}^{(1,2)} \\ \sigma_{u0,u0}^{(1,2)} & \sigma_{u0}^{(2)2} \end{bmatrix} \right)$$

$$\begin{bmatrix} e_{0j}^{(1)} \\ e_{0j}^{(2)} \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{e0}^{(1)2} & \sigma_{e0,e0}^{(1,2)} \\ \sigma_{e0,e0}^{(1,2)} & \sigma_{e0}^{(2)2} \end{bmatrix} \right)$$

Multiple response models

$$y_{ij}^{(1)} = \beta_0^{(1)} + \sum_{p=1}^P \beta_p^{(1)} x_{pij} + u_{0j}^{(1)} + e_{0ij}^{(1)}$$

$$y_{ij}^{(2)} = \beta_0^{(2)} + \sum_{p=1}^P \beta_p^{(2)} x_{pij} + u_{0j}^{(2)} + e_{0ij}^{(2)}$$

$$Var\left(y_{ij}^{(r)}\right) = \sigma_{u0}^{(r)2} + \sigma_{e0}^{(r)2}$$

$$Cov\left(y_{ij}^{(r)}, y_{i'j}^{(r)}\right) = \sigma_{u0}^{(r)2} \quad Cov\left(y_{ij}^{(r)}, y_{ij}^{(s)}\right) = \sigma_{u0,u0}^{(r,s)} + \sigma_{e0,e0}^{(r,s)}$$

$$Cov\left(y_{ij}^{(r)}, y_{i'j'}^{(r)}\right) = 0 \quad Cov\left(y_{ij}^{(r)}, y_{i'j}^{(s)}\right) = \sigma_{u0,u0}^{(r,s)}$$

Adding complexity

- All models can be adapted to include heteroscedasticity
- All models can be adapted to permit non-Normal outcome variables
 - ▲ e.g. logistic, Poisson, negative binomial, multinomial, proportional hazards...
- All the above models assume higher level residuals normally distributed
 - ▲ Convenient
 - ▲ Robust?
 - ▲ Not essential using MCMC

Fitting multilevel models

Alastair H Leyland

MRC Social and Public Health Sciences Unit
Glasgow, Scotland

Outline

- Mechanics of model fitting
 - ▲ Algorithms for IGLS/RIGLS
 - ▲ Common alternatives
- Examples and interpretation
 - ▲ Fixed part, random part, uncertainty
 - ▲ Logistic and Poisson regression models

IGLS: fixed part

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}^{(2)}\mathbf{U} + \mathbf{Z}^{(1)}\mathbf{E}$$

Stacked residuals at levels 2 and 1

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{Y}$$

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \left(\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X} \right)^{-1}$$

IGLS: random part

$$\mathbf{Y}^{**} = \text{vec}\left([\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}][\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}]^T\right)$$

$$E(\mathbf{Y}^{**}) = \mathbf{Z}^* \boldsymbol{\theta} \quad \mathbf{V}^* = \mathbf{V} \otimes \mathbf{V}$$

Stacked variances and covariances

$$\hat{\boldsymbol{\theta}} = \left(\mathbf{Z}^{*T} \mathbf{V}^{*-1} \mathbf{Z}^*\right)^{-1} \mathbf{Z}^{*T} \mathbf{V}^{*-1} \mathbf{Y}^{**}$$

$$\text{Var}(\hat{\boldsymbol{\theta}}) = \left(\mathbf{Z}^{*T} \mathbf{V}^{*-1} \mathbf{Z}^*\right)^{-1} \mathbf{Z}^{*T} \mathbf{V}^{*-1} \text{Var}(\mathbf{Y}^{**})$$

$$\mathbf{V}^{*-1} \mathbf{Z}^* \left(\mathbf{Z}^{*T} \mathbf{V}^{*-1} \mathbf{Z}^*\right)^{-1}$$

IGLS v RIGLS

IGLS estimates based on known values:

$$E\left(\left[\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right]\left[\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right]^T\right) = \mathbf{V}$$

RIGLS estimates based on estimates:

$$E\left(\left[\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right]\left[\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right]^T\right) = \\ \mathbf{V} - \mathbf{X}\left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T$$

Alternative model-fitting approaches

- IGLS/RIGLS
- EM algorithm
- Numerical integration
- MCMC – Gibbs sampling and Metropolis Hastings
- Marginal or population average models (GEE)

Residuals subject i

Equations

$bmi_i \sim N(\mu_i, \Omega)$ Standard error of mean

$bmi_i = \beta_{0i} \text{ cons}$ Random part

$\beta_{0i} = 27.463(0.064) + e_{0i}$ 95% confidence interval

$[e_{0i}] \sim N(0, \Omega_e) : \Omega_e = [26.997(0.474)]$ Variance of people have BMI within 27.5 +/- 10.2

-2 log likelihood

-2*loglikelihood(IGLS Deviance) = 39831.560(6494 of 6494 cases in use)

Name Fonts + - Add Term Estimates Nonlinear Clear Notation Responses Help

Equations

$$bmi_i \sim N(XB, \Omega)$$

$$bmi_i = \beta_{0i} \text{cons} + 1.765(0.462)a_i + -4.976(1.105)a_i^2 + -6.418(2.286)a_i^3 + -0.033(1.947)a_i^4 + 4.490(2.677)a_i^5 + -0.043(0.127)\text{male}_i$$

Polynomial in age
Gender effect

$$\beta_{0i} = 28.548(0.124) + e_{0i}$$

$$\begin{bmatrix} e_{0i} \end{bmatrix} \sim N(0, \Omega_e) : \Omega_e = \begin{bmatrix} 25.704(0.451) \end{bmatrix}$$

4.8% reduction in variance

$$-2 * \loglikelihood(IGLS \text{ Deviance}) = 39512.890(6494 \text{ of } 6494 \text{ cases in use})$$

Reduction in -2 log likelihood:
318.67, 6df

Name	Fonts	+	-	Add Term	Estimates	Nonlinear	Clear	Notation	Responses	Help
------	-------	---	---	----------	-----------	-----------	-------	----------	-----------	------

Response now at 2 levels (people in areas)

Equations

$$bmi_{ij} \sim N(XB, \Omega)$$

$$bmi_{ij} = \beta_{0ij} \text{cons} + 1.813(0.463)a_{ij} + -5.046(1.105)a_{ij}^2 + -6.698(2.288)a_{ij}^3 + 0.069(1.946)a_{ij}^4 + 4.784(2.677)a_{ij}^5 + -0.038(0.126)\text{male}_{ij}$$

$$\beta_{0ij} = 28.542(0.128) + u_{0ij} + e_{0ij}$$

Random part includes area effects

$$\begin{bmatrix} u_{0ij} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 0.389(0.131) \end{bmatrix}$$

95% of 0.389 is within 9.86

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim N(0, \Omega_e) : \Omega_e = \begin{bmatrix} 25.318(0.455) \end{bmatrix}$$

95% of 25.318 is within 9.86

- of highest order as

$$-2 * \loglikelihood(\text{IGLS Deviance}) = 39500.890(6494 \text{ of } 6494 \text{ cases in use})$$

Reduction in -2 log likelihood
12.00, 1df

Name	Fonts	+	-	Add Term	Estimates	Nonlinear	Clear	Notation	Responses	Help
------	-------	---	---	----------	-----------	-----------	-------	----------	-----------	------

Equations

$$bmi_{ij} \sim N(XB, \Omega)$$

$$bmi_{ij} = \beta_{0ij} \text{cons} + 1.555(0.486)a_{ij} + -4.776(1.210)a^2_{ij} + -6.418(2.342)a^3_{ij} + -0.057(2.026)a^4_{ij} +$$

$$4.606(2.745)a^5_{ij} + -0.038(0.126)\text{male}_{ij} + 0.845(0.295)\text{educ}<15_{ij} + 0.979(0.195)\text{educ}15_{ij} +$$

$$0.931(0.184)\text{educ}16_{ij} + 0.459(0.205)\text{educ}17-18_{ij}$$

$$\beta_{0ij} = 27.823(0.184) + u_{0ij} + e_{0ij}$$

Dummy variables for education included

$$[u_{0ij}] \sim N(0, \Omega_u) : \Omega_u = [0.296(0.123)] \text{ Area variance 23.9\% lower}$$

$$[e_{0ij}] \sim N(0, \Omega_e) : \Omega_e = [25.256(0.454)] \text{ Individual variance 0.2\% lower}$$

-2*loglikelihood(IGLS Deviance) = 39454.460 (6492 of 6494 cases in use)

Education missing for two people

Name	Fonts	+	-	Add Term	Estimates	Nonlinear	Clear	Notation	Responses	Help
------	-------	---	---	----------	-----------	-----------	-------	----------	-----------	------

Equations

$$bmi_{ij} \sim N(XB, \Omega)$$

$$bmi_{ij} = \beta_{0ij} \text{cons} + 1.550(0.486)a_{ij} + -4.787(1.209)a^2_{ij} + -6.128(2.340)a^3_{ij} + -0.171(2.024)a^4_{ij} +$$

$$4.260(2.743)a^5_{ij} + -0.030(0.126)male_{ij} + 0.723(0.297)educ<15_{ij} + 0.851(0.197)educ15_{ij} +$$

$$0.845(0.186)educ16_{ij} + 0.431(0.203)educ17-18_{ij} + 0.501(0.210)carst2_j +$$

$$0.779(0.215)carst3_j + 0.792(0.228)carst4_j + 0.697(0.224)carst5_j$$

$$\beta_{0ij} = 27.362(0.219) + u_{0ij} + e_{0ij}$$

Variables added at area level

$$[u_{0ij}] \sim N(0, \Omega_u) : \Omega_u = [0.229(0.117)] \text{Area variance 22.6\% lower}$$

$$[e_{0ij}] \sim N(0, \Omega_e) : \Omega_e = [25.243(0.454)] \text{Individual variance unchanged}$$

$-2 * \loglikelihood(IGLS \text{ Deviance}) = 39437.070(6492 \text{ of } 6494 \text{ cases in use})$

Name	Fonts	+	-	Add Term	Estimates	Nonlinear	Clear	Notation	Responses	Help
------	-------	---	---	----------	-----------	-----------	-------	----------	-----------	------

3 levels (people in households within areas)

Equations

$$\text{bmi}_{ijk} \sim N(XB, \Omega)$$

$$\text{bmi}_{ijk} = \beta_{0ijk} \text{cons} + 1.482(0.502)a_{ijk} + -5.115(1.233)a^2_{ijk} + -5.972(2.396)a^3_{ijk} +$$

$$0.369(2.047)a^4_{ijk} + 4.449(2.787)a^5_{ijk} + 0.009(0.118)\text{male}_{ijk} + 0.681(0.299)\text{educ}<15_{ijk} +$$

$$0.762(0.200)\text{educ}15_{ijk} + 0.825(0.188)\text{educ}16_{ijk} + 0.369(0.203)\text{educ}17-18_{ijk} +$$

$$0.495(0.212)\text{carst}2_k + 0.772(0.217)\text{carst}3_k + 0.784(0.230)\text{carst}4_k + 0.692(0.227)\text{carst}5_k$$

$$\beta_{0ijk} = 27.418(0.222) + v_{0k} + u_{0jk} + e_{0ijk}$$

$$\begin{bmatrix} v_{0k} \end{bmatrix} \sim N(0, \Omega_v) : \Omega_v = \begin{bmatrix} 0.078(0.121) \end{bmatrix}$$

$$\begin{bmatrix} u_{0jk} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 5.168(0.542) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ijk} \end{bmatrix} \sim N(0, \Omega_e) : \Omega_e = \begin{bmatrix} 20.275(0.569) \end{bmatrix}$$

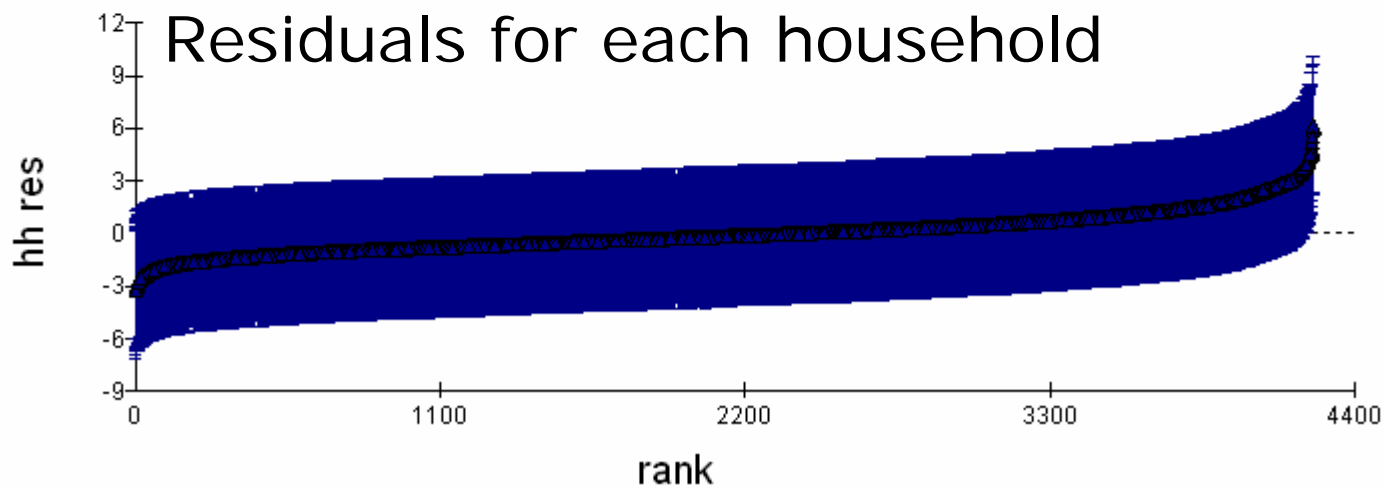
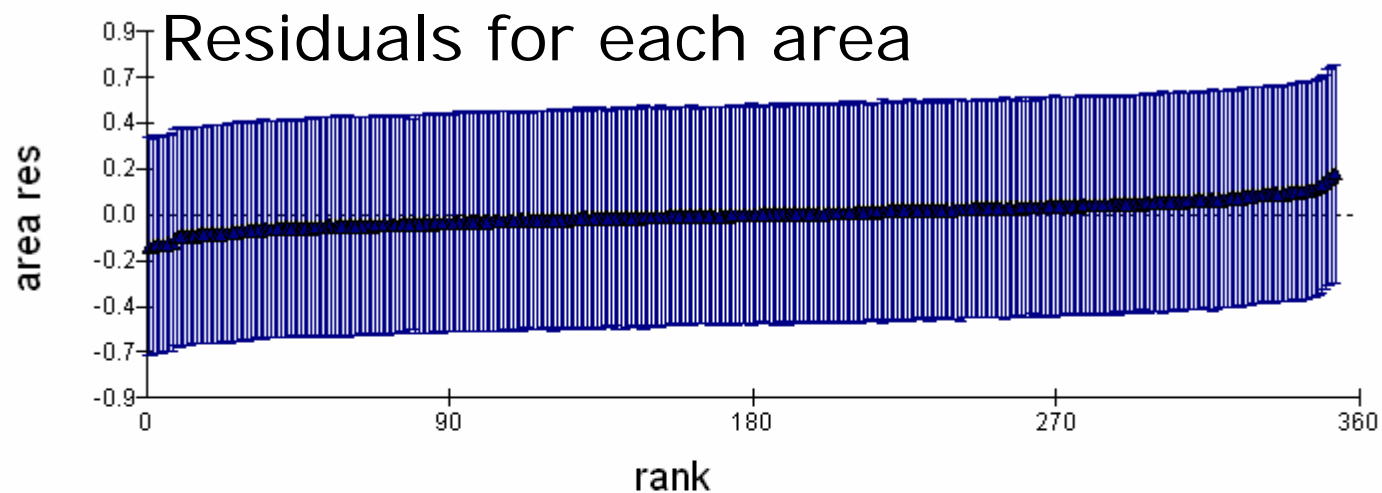
Total variance unchanged
20.2% at household level

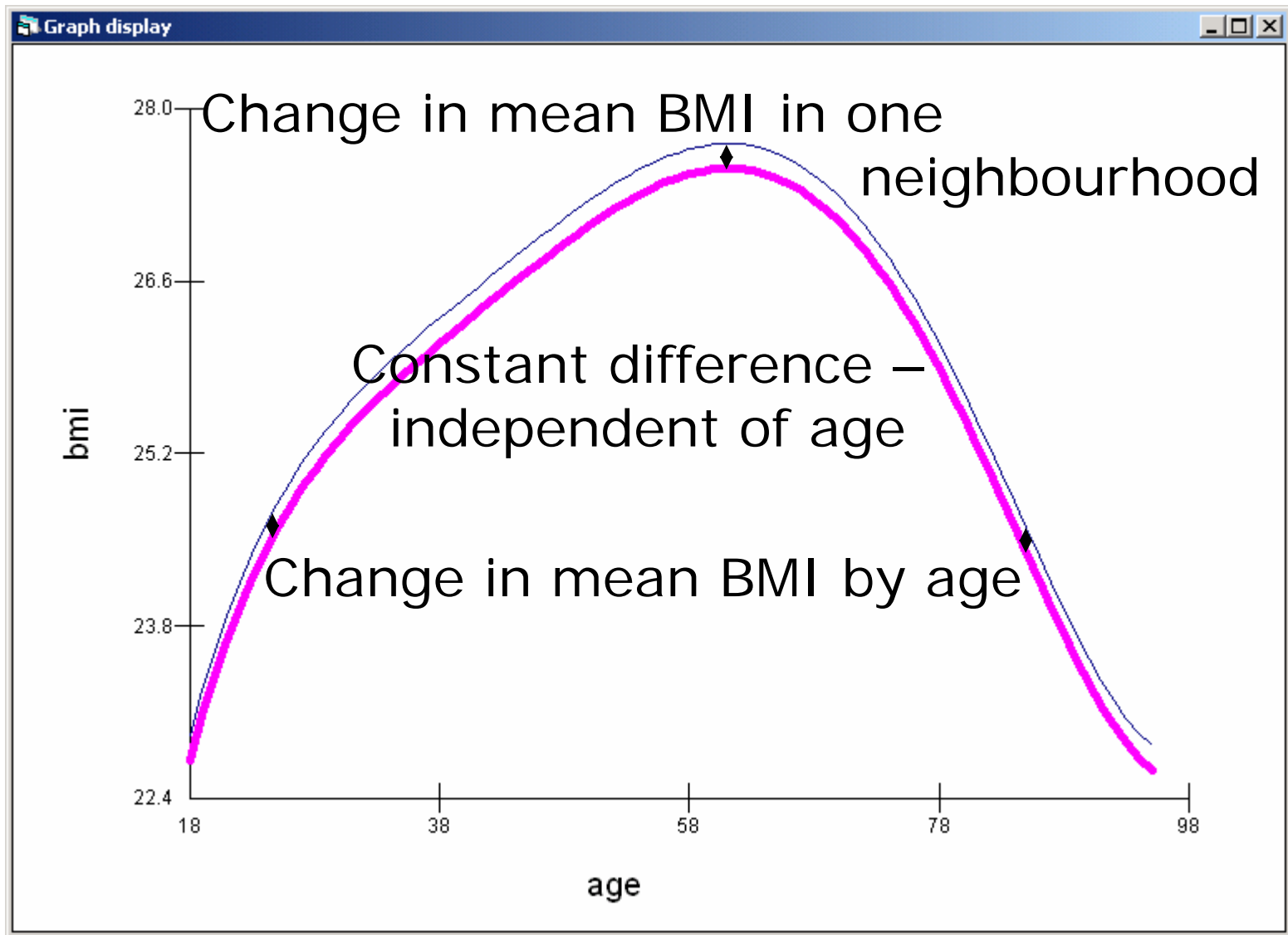
$$-2*\text{loglikelihood(IGLS Deviance)} = 39350.100(6492 \text{ of } 6494 \text{ cases in use})$$

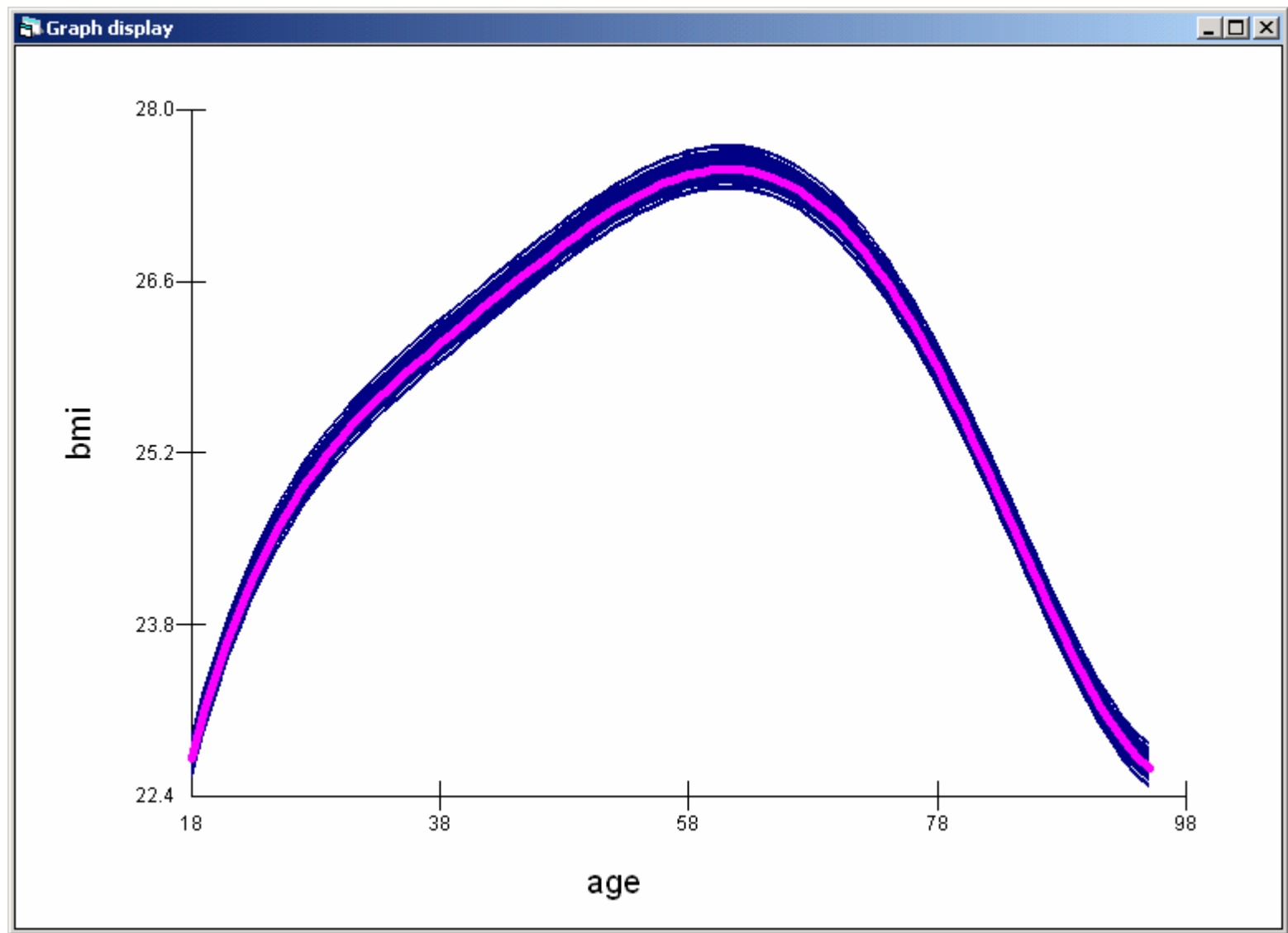
Name	Fonts	+	-	Add Term	Estimates	Nonlinear	Clear	Notation	Responses	Help
------	-------	---	---	----------	-----------	-----------	-------	----------	-----------	------

Reduction in -2 log likelihood
86.97, 1df

Graph display







Equations

$bmi_{ijk} \sim N(XB, \Omega)$

Intercept no longer random

$bmi_{ijk} = 27.508(0.220)cons + 1.201(0.488)a_{ijk} + -5.344(1.198)a^2_{ijk} + -5.285(2.327)a^3_{ijk} +$
 $0.885(1.978)a^4_{ijk} + 4.381(2.696)a^5_{ijk} + \beta_{6ijk}male_{ijk} + 0.652(0.290)educ<15_{ijk} +$
 $0.719(0.194)educ15_{ijk} + 0.811(0.182)educ16_{ijk} + 0.386(0.198)educ17-18_{ijk} +$
 $0.470(0.207)carst2_k + 0.645(0.212)carst3_k + 0.652(0.224)carst4_k + 0.573(0.221)carst5_k$
 $+ e_{15ijk}female_{ijk} + u_{15ijk}female_{ijk} + v_{15k}female_{ijk}$

Male random at all levels

Female random at all levels

$\beta_{6ijk} = 0.021(0.114) + v_{6k} + u_{6jk} + e_{6ijk}$

$\begin{bmatrix} v_{6k} \\ v_{15k} \end{bmatrix} \sim N(0, \Omega_v) : \Omega_v = \begin{bmatrix} 0.081(0.171) & 0.084(0.143) \\ 0.084(0.143) & 0.088(0.229) \end{bmatrix}$

High correlation at area level = 0.99

$\begin{bmatrix} u_{6jk} \\ u_{15jk} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 0.829(1.447) & 5.178(0.538) \\ 5.178(0.538) & 7.424(2.209) \end{bmatrix}$

High correlation at hh level = 2.1!!

$\begin{bmatrix} e_{6ijk} \\ e_{15ijk} \end{bmatrix} \sim N(0, \Omega_e) : \Omega_e = \begin{bmatrix} 18.350(1.508) & 0 \\ 0 & 23.112(2.208) \end{bmatrix}$

No correlation between male and female at level

More variation within households for females than males

$-2*loglikelihood(IGLS Deviance) = 39174.750(6492 \text{ of } 6494 \text{ cases in use})$

Name Fonts + - Add Term Estimates Nonlinear Clear Notation Responses Help

Equations

$$\text{bmi}_{ijk} \sim N(XB, \Omega)$$

$$\text{bmi}_{ijk} = 27.513(0.222)\text{cons} + 1.196(0.492)a_{ijk} + -5.361(1.188)a^2_{ijk} + -5.254(2.341)a^3_{ijk} + 0.963(1.959)a^4_{ijk} + 4.386(2.702)a^5_{ijk} + \beta_{6ijk}\text{male}_{ijk} + 0.662(0.293)\text{educ}<15_{ijk} + 0.722(0.195)\text{educ}15_{ijk} + 0.805(0.186)\text{educ}16_{ijk} + 0.385(0.200)\text{educ}17-18_{ijk} + 0.468(0.207)\text{carst}2_k + 0.647(0.214)\text{carst}3_k + 0.658(0.225)\text{carst}4_k + 0.561(0.220)\text{carst}5_k + e_{15ijk}\text{female}_{ijk} + u_{15jk}\text{female}_{ijk} + v_{15k}\text{female}_{ijk}$$

$$\beta_{6ijk} = 0.009(0.116) + v_{6k} + u_{6jk} + e_{6ijk}$$

$$\begin{bmatrix} v_{6k} \\ v_{15k} \end{bmatrix} \sim N(0, \Omega_v) : \Omega_v = \begin{bmatrix} 0.112(0.070) & 0.119(0.074) \\ 0.119(0.074) & 0.130(0.082) \end{bmatrix} \text{ High correlation at area level } = 0.99$$

$$\begin{bmatrix} u_{6jk} \\ u_{15jk} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 2.674(0.651) & 4.569(0.533) \\ 4.569(0.533) & 10.381(1.542) \end{bmatrix} \text{ High correlation at hh level } = 0.87$$

$$\begin{bmatrix} e_{6ijk} \\ e_{15ijk} \end{bmatrix} \sim N(0, \Omega_e) : \Omega_e = \begin{bmatrix} 16.537(0.725) & 20.258(1.534) \\ 0 & 0 \end{bmatrix}$$

MCMC deviance (and DIC) instead of -2 log likelihood

$$\text{Deviance}(MCMC) = 37345.660 \text{ (6492 of 6494 cases in use)}$$

Name	Fonts	+	-	Add Term	Estimates	Nonlinear	Clear	Notation	Responses	Help
------	-------	---	---	----------	-----------	-----------	-------	----------	-----------	------

Back to 2-level model: people in areas

Equations

Random slope with age

$$bmi_{ij} \sim N(XB, \Omega)$$

$$bmi_{ij} = \beta_{0ij} \text{cons} + \beta_{1j} a_{ij} + -4.788(1.210)a_{ij}^2 + -6.104(2.344)a_{ij}^3 + -0.115(2.028)a_{ij}^4 +$$

$$4.263(2.749)a_{ij}^5 + -0.028(0.126)male_{ij} + 0.720(0.297)educ<15_{ij} + 0.848(0.197)educ15_{ij} +$$

$$0.843(0.186)educ16_{ij} + 0.423(0.203)educ17-18_{ij} + 0.500(0.210)carst2_j +$$

$$0.771(0.215)carst3_j + 0.792(0.228)carst4_j + 0.698(0.224)carst5_j$$

$$\beta_{0ij} = 27.366(0.219) + u_{0j} + e_{0ij}$$

$$\beta_{1j} = 1.549(0.487) + u_{1j}$$

Variance between areas

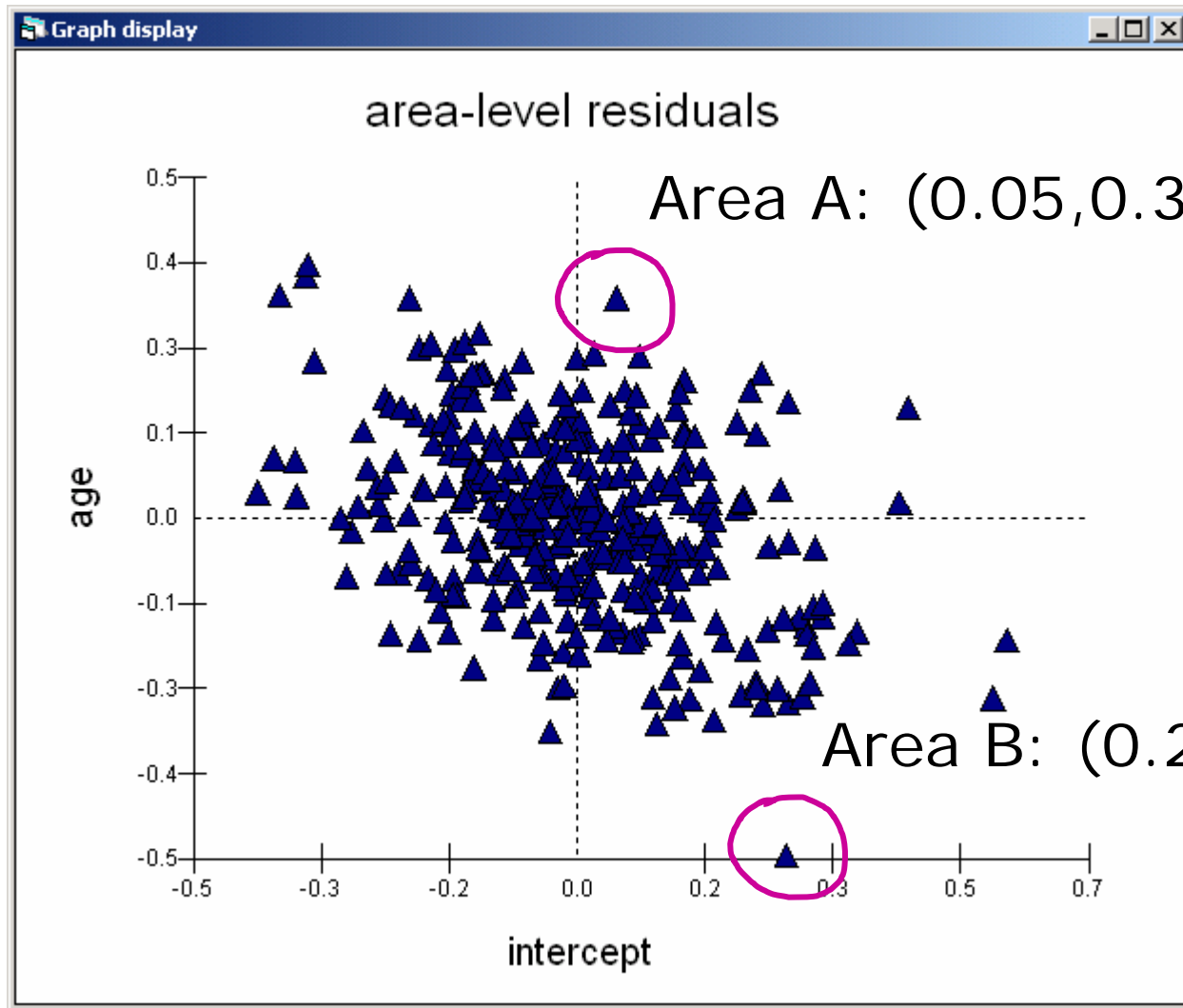
$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 0.214(0.134) & -0.001(0.203) \\ -0.001(0.203) & 0.362(0.566) \end{bmatrix} \text{ in age effect}$$

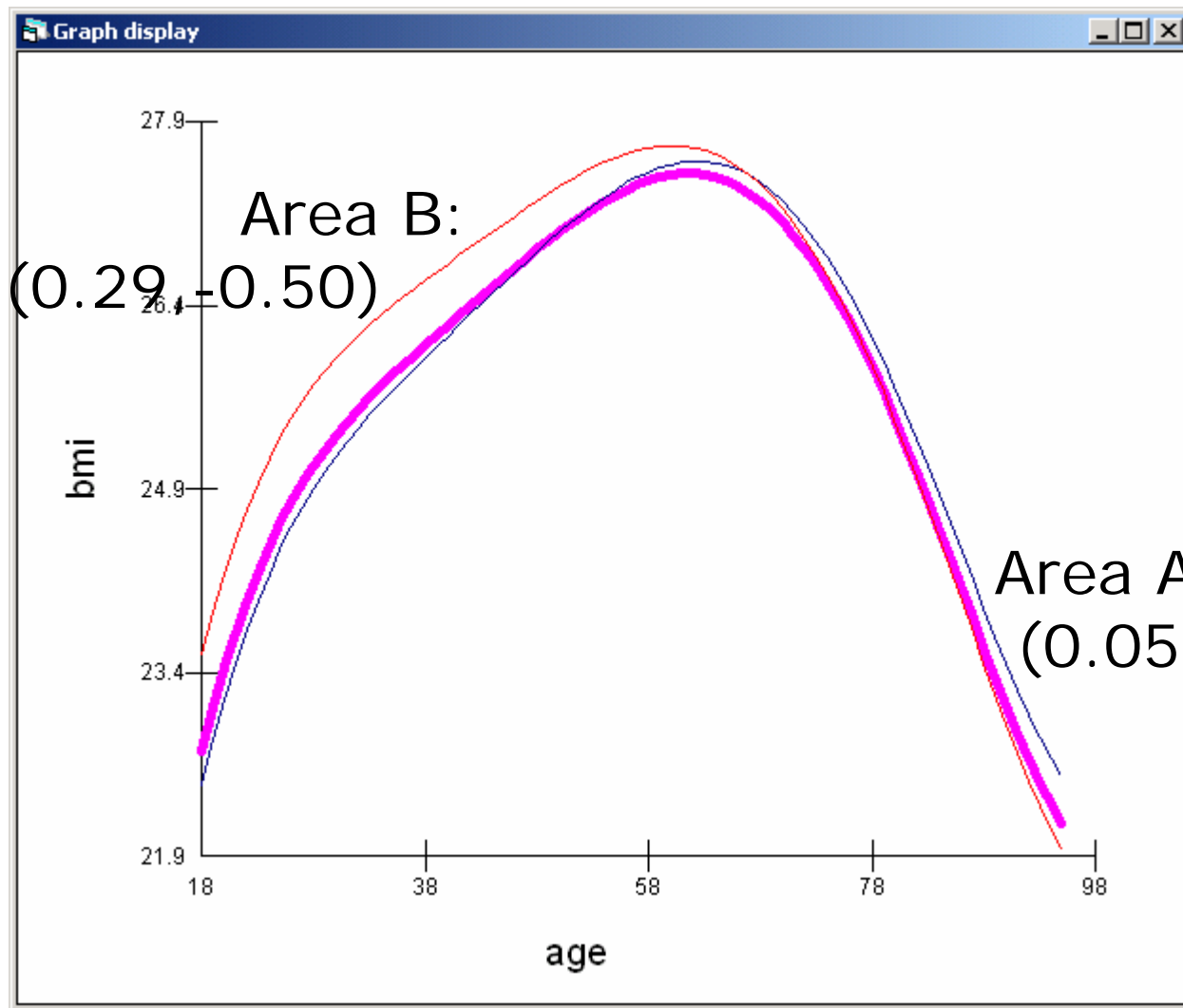
No correlation between intercept

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim N(0, \Omega_e) : \Omega_e = \begin{bmatrix} 25.181(0.462) \end{bmatrix} \text{ and slope}$$

$-2 * \loglikelihood(IGLS \text{ Deviance}) = 39436.500(6492 \text{ of } 6494 \text{ cases in use})$

Name	Fonts	+	-	Add Term	Estimates	Nonlinear	Clear	Notation	Responses	Help
------	-------	---	---	----------	-----------	-----------	-------	----------	-----------	------





Logan Binomial distribution customized to >30

Equations

obese_{ij} ~ Binomial(cons_{ij}, π_{ij})

logit(π_{ij}) = β_{0j}cons + 0.685(0.225)a_{ij} + -0.804(0.652)a²_{ij} + -2.365(1.197)a³_{ij} + -1.730(1.321)a⁴_{ij} +
 0.591(1.662)a⁵_{ij} + -0.170(0.059)male_{ij} + 0.502(0.136)educ<15_{ij} +
 0.418(0.093)educ15_{ij} + 0.399(0.090)educ16_{ij} + 0.222(0.101)educ17-18_{ij} +
 0.201(0.098)carst2_j + 0.407(0.099)carst3_j + 0.442(0.104)carst4_j + 0.392(0.103)carst5_j

β_{0j} = -1.324(0.106) + u_{0j} Param 0.184 (95% CIs: -0.751 to 0.0015)

[u_{0j}] ~ N(0, Ω_u) : Ω_u = [0.031(0.023)] ratios

var(obese_{ij} | π_{ij}) = π_{ij}(1 - π_{ij}) / cons_{ij} DE associated (only 19-3528) scale
 relative 0.1002.5 centile = 1.99

Variance at lowest level related to probability
 Pseudo-likelihood so no -2 log likelihood
 of outcome

Home Fonts + - Add Term Estimates Nonlinear Clear Notation Responses Help

Summary

- Standard interpretation of fixed part
- Random part refers to variation
- Variance can be partitioned
- Variance can be explained
- Random intercepts – parallel lines
- Random slopes – relationship varies across contexts
- ML GLMs require different interpretations of variances as well as fixed part

Investigating sparse clusters: clustering within households

Alastair H Leyland

MRC Social and Public Health Sciences Unit
Glasgow, Scotland

Outline

- How many units at one level within units at the next to make multilevel modelling worthwhile?
- Typically few individuals per household
- Household frequently omitted as a level
- Analyses stratified by sex particularly likely to omit household
- ...but does it make any difference?

Data

- 2003 Scottish Health Survey
- 7901 adults aged 18-95
- Clustered in 5063 households, 356 postcode sectors
- 2512 (49.6%) households have 1 adult
- 245 (3.1%) households have 3+ adults

Outcomes

- Continuous
 - ▲ BMI
 - ▲ Systolic blood pressure
- Dichotomous
 - ▲ Smoking
 - ▲ Consultation with GP
 - ▲ Oily fish at least once per week
 - ▲ Five+ fruit & veg per day
- Adjust for age, sex, individual social class, education, area deprivation

Strategy

- Comparisons of models including and excluding household
- What is the appropriate model to use to make comparisons?
- How should sex effects be included?
- Compare different models using DIC

Model A

Additive effect for sex

$$y_{ijk} = \beta_0 m_{ijk} + \beta_1 f_{ijk} + \sum_{p=2}^P \beta_p x_{pijk} + v_{0k} + u_{0jk} + e_{0ijk}$$

$$v_{0k} \sim N(0, \sigma_{v0}^2) \quad u_{0jk} \sim N(0, \sigma_{u0}^2)$$

$$e_{0ijk} \sim N(0, \sigma_{e0}^2)$$

Model B

Fixed part interaction with sex

$$y_{ijk} = \beta_0 m_{ijk} + \beta_1 f_{ijk} + \sum_{p=2}^P \beta_p x_{pijk} m_{ijk} + \sum_{p=P+1}^{2P-1} \beta_p x_{pijk} f_{ijk}$$

$$+ v_{0k} + u_{0jk} + e_{0ijk}$$

$$v_{0k} \sim N(0, \sigma_{v0}^2) \quad u_{0jk} \sim N(0, \sigma_{u0}^2)$$

$$e_{0ijk} \sim N(0, \sigma_{e0}^2)$$

Model C

Random part interaction with sex

$$y_{ijk} = \beta_0 m_{ijk} + \beta_1 f_{ijk} + \sum_{p=2}^P \beta_p x_{pijk} \\ + \left(v_{0k} + u_{0jk} + e_{0ijk} \right) m_{ijk} + \left(v_{1k} + u_{1jk} + e_{1ijk} \right) f_{ijk} \\ \begin{bmatrix} v_{0k} \\ v_{1k} \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{v0}^2 & \sigma_{v01} \\ \sigma_{v01} & \sigma_{v1}^2 \end{bmatrix} \right) \begin{bmatrix} u_{0jk} \\ u_{1jk} \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{u0}^2 & \sigma_{u01} \\ \sigma_{u01} & \sigma_{u1}^2 \end{bmatrix} \right) \\ \begin{bmatrix} e_{0ijk} \\ e_{1ijk} \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{e0}^2 & 0 \\ 0 & \sigma_{e1}^2 \end{bmatrix} \right)$$

Model D

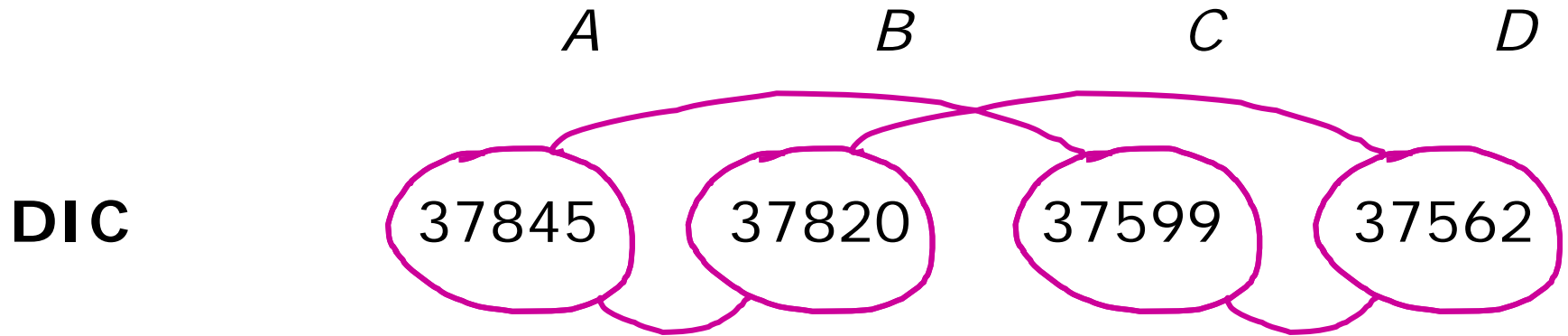
Fixed and random part interactions with sex

$$y_{ijk} = \beta_0 m_{ijk} + \beta_1 f_{ijk} + \sum_{p=2}^P \beta_p x_{pijk} m_{ijk} + \sum_{p=P+1}^{2P-1} \beta_p x_{pijk} f_{ijk} \\ + \left(v_{0k} + u_{0jk} + e_{0ijk} \right) m_{ijk} + \left(v_{1k} + u_{1jk} + e_{1ijk} \right) f_{ijk}$$

$$\begin{bmatrix} v_{0k} \\ v_{1k} \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{v0}^2 & \sigma_{v01} \\ \sigma_{v01} & \sigma_{v1}^2 \end{bmatrix} \right) \begin{bmatrix} u_{0jk} \\ u_{1jk} \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{u0}^2 & \sigma_{u01} \\ \sigma_{u01} & \sigma_{u1}^2 \end{bmatrix} \right)$$

$$\begin{bmatrix} e_{0ijk} \\ e_{1ijk} \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{e0}^2 & 0 \\ 0 & \sigma_{e1}^2 \end{bmatrix} \right)$$

Model comparison for BMI



Effect of fixed part interactions

Model D*

Full interactions with sex, no household effect

$$y_{ijk} = \beta_0 m_{ijk} + \beta_1 f_{ijk} + \sum_{p=2}^P \beta_p x_{pijk} m_{ijk} + \sum_{p=P+1}^{2P-1} \beta_p x_{pijk} f_{ijk} \\ + \left(v_{0k} + e_{0ijk} \right) m_{ijk} + \left(v_{1k} + e_{1ijk} \right) f_{ijk}$$

$$\begin{bmatrix} v_{0k} \\ v_{1k} \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{v0}^2 & \sigma_{v01} \\ \sigma_{v01} & \sigma_{v1}^2 \end{bmatrix} \right)$$

$$\begin{bmatrix} e_{0ijk} \\ e_{1ijk} \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{e0}^2 & 0 \\ 0 & \sigma_{e1}^2 \end{bmatrix} \right)$$

Model comparison for BMI

	D	D^*
DIC	37562	37963

BMI : effect on random part

	D		D*	
σ_{vM}^2	0.11	(0.00-0.37)	0.18	(0.00-0.52)
σ_{vMF}	0.02	(-0.05-0.17)	0.12	(-0.03-0.39)
σ_{vF}^2	0.09	(0.00-0.45)	0.23	(0.00-0.69)
σ_{uM}^2	3.04	(1.62-5.00)	-	-
σ_{uMF}	4.67	(3.49-5.66)	-	-
σ_{uF}^2	8.73	(5.09-12.93)	-	-
σ_{eM}^2	16.08	(14.15-17.86)	19.03	(18.04-20.06)
σ_{eF}^2	21.34	(17.30-25.15)	29.84	(28.43-31.33)

BMI : effect on random part

	D		D*	
$\% (M, area)$	0.6	(0.0-1.9)	0.9	(0.0-2.7)
$\rho (area)$	0.27	(-0.80-0.96)	0.53	(-0.78-0.99)
$\% (F, area)$	0.3	(0.0-1.5)	0.8	(0.0-2.3)
$\% (M, hh)$	15.8	(8.4-25.7)	-	-
$\rho (hh)$	0.94	(0.77-1.00)	-	-
$\% (F, hh)$	28.9	(17.0-42.4)	-	-
$\% (M, indiv)$	83.6	(73.7-91.0)	99.1	(97.3-100.0)
$\% (F, indiv)$	70.8	(57.4-82.7)	99.3	(97.7-100.0)

BMI : effect on fixed part

Age on leaving education

	D		D*		% diff	
M <15 yrs	0.55	(0.39)	0.52	(0.40)	-4.69	(1.07)
M 15 yrs	0.59	(0.28)	0.62	(0.28)	5.47	(0.68)
M 16 yrs	0.71	(0.25)	0.67	(0.26)	-4.40	(1.46)
M 17-18 yrs	0.37	(0.27)	0.40	(0.28)	6.28	(1.31)
F <15 yrs	0.31	(0.47)	0.35	(0.48)	10.51	(1.70)
F 15 yrs	0.59	(0.32)	0.65	(0.32)	10.90	(1.42)
F 16 yrs	0.59	(0.29)	0.61	(0.30)	3.37	(1.43)
F 17-18 yrs	0.07	(0.30)	0.16	(0.30)	114.1	(0.87)

Relative to those leaving at 19 years onwards

Model comparisons

	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>
BMI	37845	37820	37599	37562
SBP	42609	42452	42513	42210
GP	7322	7325	7107	6943
Smoke	7550	7539	7530	7469
Fish	4432	4324	4273 [#]	4267 [#]
Fr&Veg	6930	6897	6710	6518

[#] Sex interaction in random part at area level only

Model comparisons

	D	D^*
BMI	37562	37963
SBP	42210	42505
GP	6943	7353
Smoke	7469	8133
Fish	4267 [#]	7652 [#]
Fr&Veg	6518	7643

[#] Sex interaction in random part at area level only

Effect on random part % variance at household level

	M		F	
BMI	15.8	(8.4-25.7)	28.9	(17.0-42.4)
SBP	32.2	(10.6-53.0)	12.1	(2.8-26.0)
GP	38.0	(0.0-77.6)	46.7	(23.8-67.8)
Smoke	41.4	(25.1-53.2)	59.4	(28.6-74.2)
Fish	86.4	(82.5-89.9)	87.3	(83.8-90.4)
Fr&Veg	68.2	(49.2-81.7)	63.0	(31.0-82.7)

Effect on fixed part: Range of % bias

	<i>Est</i>	<i>SE</i>
BMI	(-8.9-114.1)	(0.3-2.0)
SBP	(-169.3-16.4)	(-0.6-2.0)
GP	(-207.2--28.6)	(-81.4--30.1)
Smoke	(-48.3--10.1)	(-71.6--23.2)
Fish	(-128.0--52.0)	(-64.5--59.5)
Fr&Veg	(-66.4--35.0)	(-66.4--49.4)

Link between cluster-specific and population average estimates

$$\beta^* = \beta / \sqrt{1 + \frac{768 \times \sigma_u^2}{225 \times \pi^2}} \approx \beta / \sqrt{1 + 0.346 \times \sigma_u^2}$$

Male variances and divisors (relating CS and PA estimates)

	D	Divisor	D^*	Divisor
GP	3.617	1.500	0.033	1.006
Smoke	2.419	1.355	0.040	1.007
Fish	28.883	3.315	0.316	1.053
Fr&Veg	10.367	2.141	0.257	1.044

Conclusions

- Even when few individuals per cluster, effect of clustering may be marked
- Health behaviours – including health seeking – strongly clustered within households
- Even when stratified by sex high correlation within households (and areas)
- Substantial impact on variances (study design, power)
- Impact on precision of fixed parameters may be less marked

An applied example: The Victorian Lifestyle and Neighbourhood Environment Study (VicLANES)

A/Prof Anne Kavanagh
Key Centre for Women's Health in Society
School of Population Health
University of Melbourne



Why place matters



Why multi-level analysis?

- **Data is often organised hierarchically such as people within postcodes, pupils in schools or classrooms, patients with in hospitals or wards, cancer patients treated by different surgeons.**
- **Differences between higher-level units are noted (eg mortality rates, school performance, adverse event rates in hospitals or with different surgeons etc)**

Why multi-level analysis?

- **But the question as to whether places, schools, hospitals or surgeons matter cannot be straightforwardly answered because we know that different types of people are in different places, schools, hospitals or see different surgeons**
- **Is it context (the effect of a place, school, hospital) or composition (effects due to different types of people) that explain differences between higher-level units?**

What is multi-level analysis?

- **Takes into account the hierarchical structure of data (eg people nested in postcodes within states).**
- **Potentially avoids atomistic and ecological fallacy**
- **Describes the effects of fixed effect variables operating at the multiple scales (usually geographical).**
- **Enables the partitioning of variation/variance between levels.**
- **Enables exploration of variation in variables between higher level units (contextual heterogeneity)**

What is a multi-level model?

Variance components

$$Y_{ij} = \beta_{0j} + R_{ij}$$

$$\beta_{0j} = \lambda_{00} + U_{0j}$$

$$Y_{ij} = \lambda_{00} + U_{0j} + R_{ij}$$

Random intercepts

$$Y_{ij} = \gamma_{00} + \beta_1 X_{ij} + \beta_1 Z_j + U_{0j} + R_{ij}$$

What is a multi-level model?

Consider
$$Y_{ij} = \gamma_{00} + \beta_1 X_{ij} + U_{0j} + R_{ij}$$

And not only can the intercept vary but the slope can as well (random slopes)

$$\beta_{0j} = \gamma_{00} + U_{0j} \text{ \& \ } \beta_{1j} = \gamma_{10} + U_{1j}$$

$$Y_{ij} = \gamma_{00} + \beta_1 X_{ij} + U_{0j} + U_{1j} X_{ij} + R_{ij}$$

Fixed part

$$\gamma_{00} + \beta_1 X_{ij} \qquad U_{0j} + U_{1j} X_{ij} + R_{ij}$$

$U_{1j} X_{ij}$ Denotes a random interaction between higher level unit and X

Partitioning variance

- Normally distributed outcomes:
 - Intraclass correlation coefficient = $\frac{U_{0j}}{U_{0j} + R_{ij}}$

Partitioning variance logistic model

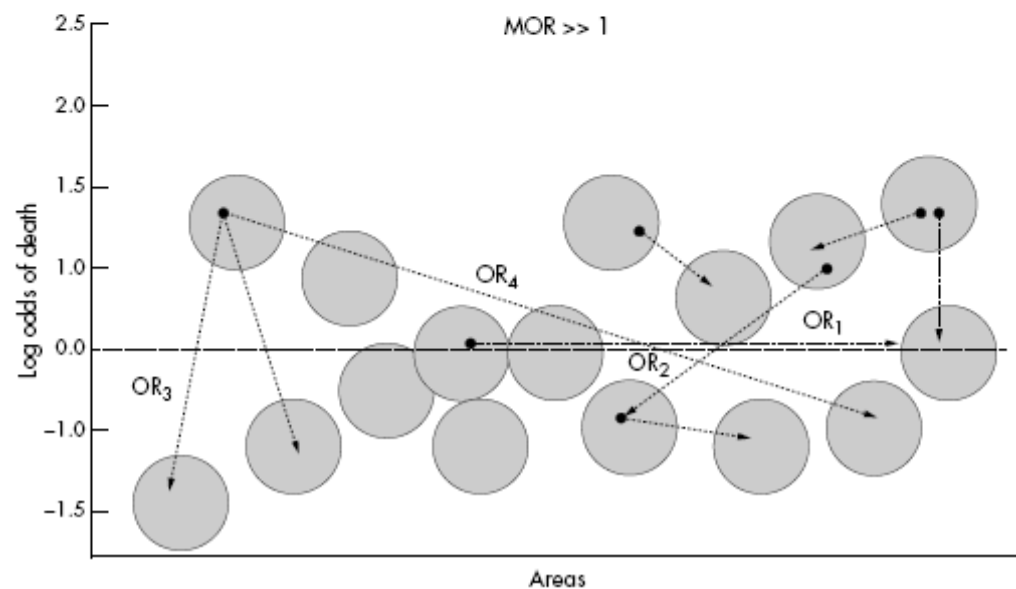
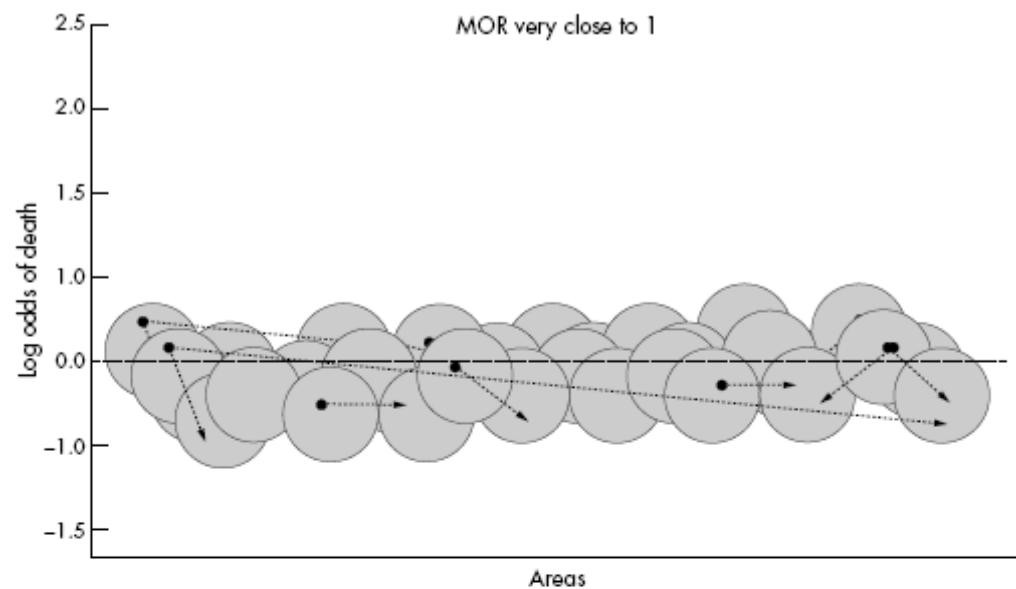
- Logistic outcomes
 - Several potential methods:
 - Methods that convert the individual and area level variance to same scale e.g.

$$ICC = V_A / (V_A + 3.29) \quad (5)$$

- Median odds ratio: the median value of the odds ratio between the area at highest risk and the area at lowest risk

$$\begin{aligned} MOR &= \exp[\sqrt{(2 \times V_A)} \times 0.6745] \\ &\approx \exp(0.95\sqrt{V_A}) \end{aligned} \quad (6)$$

- Interval odds ratio: takes into account the area level variance when interpreting area level fixed effects. 80% interval odds ratio. Usual odds ratio compares mean odds in contrasting areas. IOR takes into specific area-level residuals when comparing persons in contrasting (e.g. high and low SES) areas. If the 80% interval is wide and includes 1 then the fixed effect (e.g. area SES) is not important in explaining area level differences.



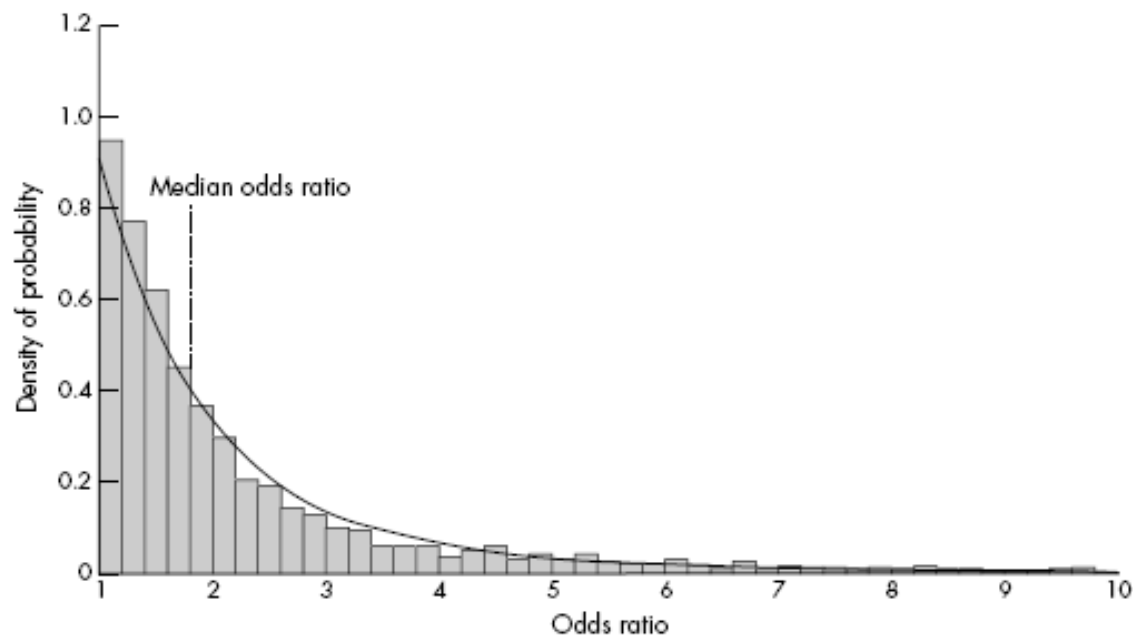
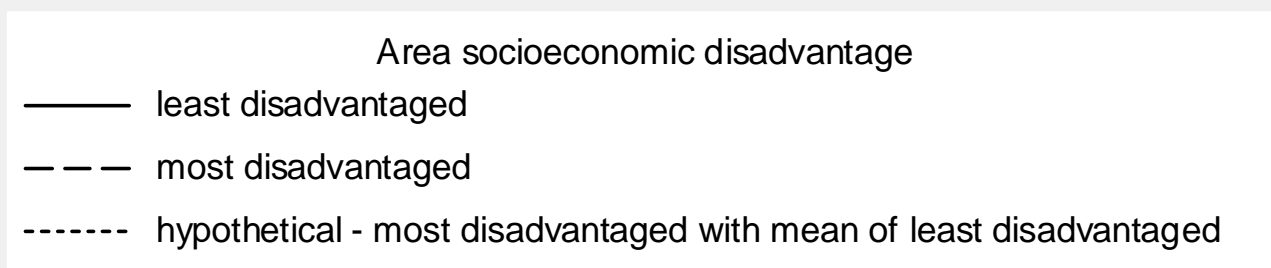
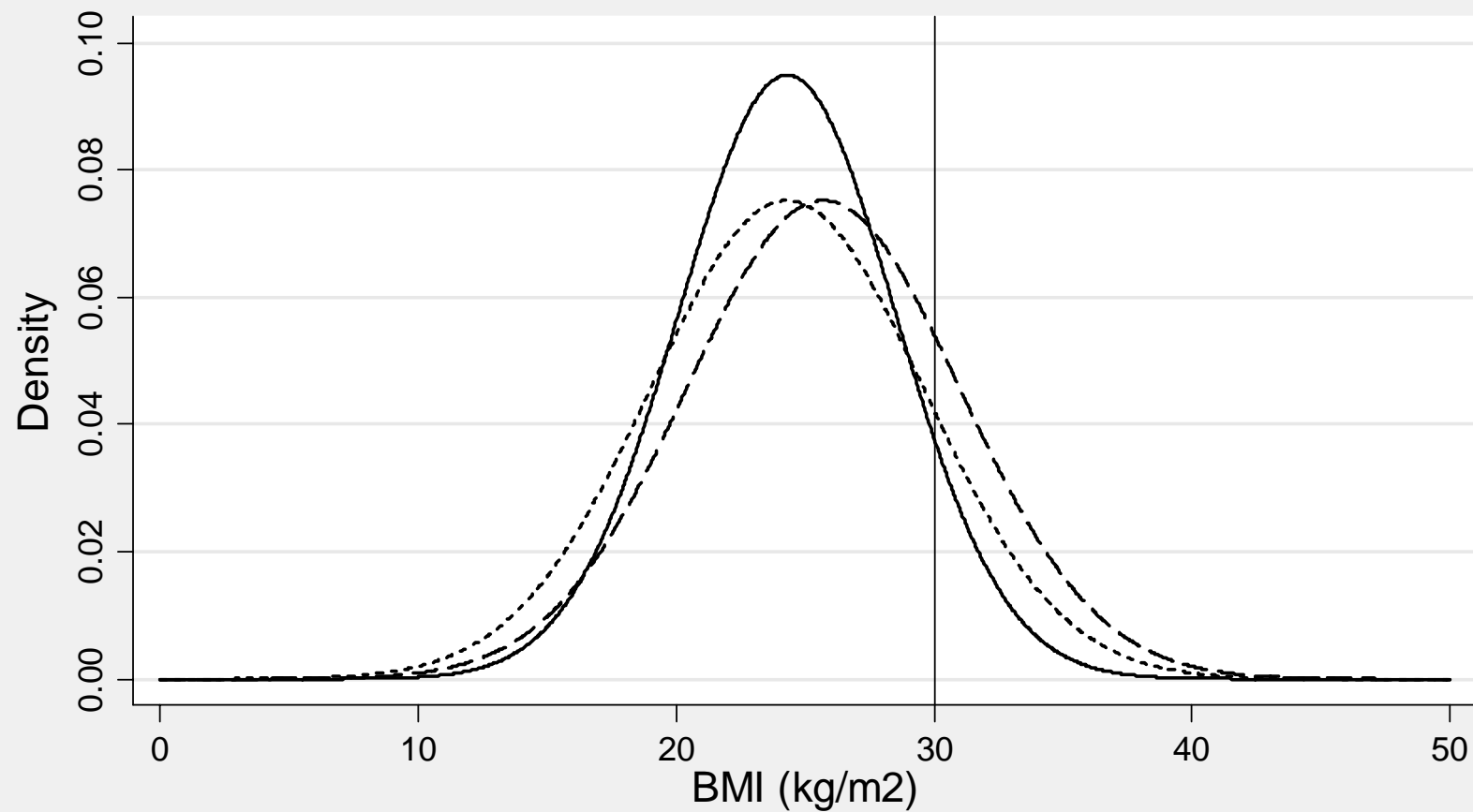


Figure 2 Considering the area level residuals of the multilevel model, we computed the odds ratio between the person at lowest risk and the individual at highest risk for each pair of persons from different areas. We present the distribution of this odds ratio for the 56 million pairs of persons from different areas that can be formed in our sample of 10 723 people. As shown in the figure, the MOR is defined as the median value of the distribution. Practically the MOR is very easy to calculate (see formula 6 in the text).



Source data taken from King et al (2005)

Rationale for VicLANES

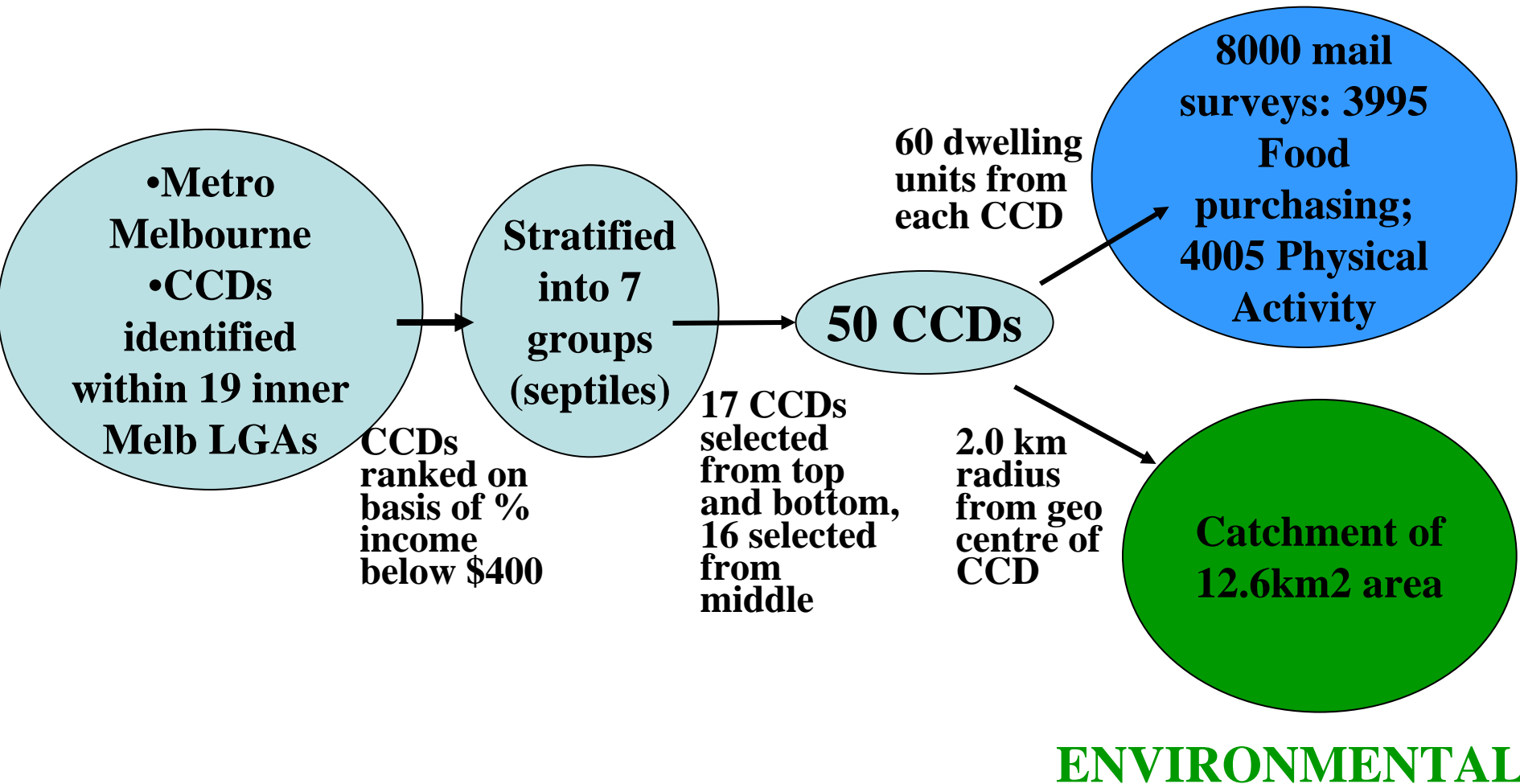
- Diet and physical activity are major contributors to the overall burden of disease
- Physical activity, diet and food-purchasing are often conducted within local areas
- If environments influence what people eat and how much they exercise then there is potential to think about ways to make environments more health promoting

Why is VicLANES unique?

- Most studies collect information about environment and individuals separately
- VicLANES links environmental data and individual data to work out the contribution of environmental and individual variables to health behaviours

Sampling plan

INDIVIDUAL

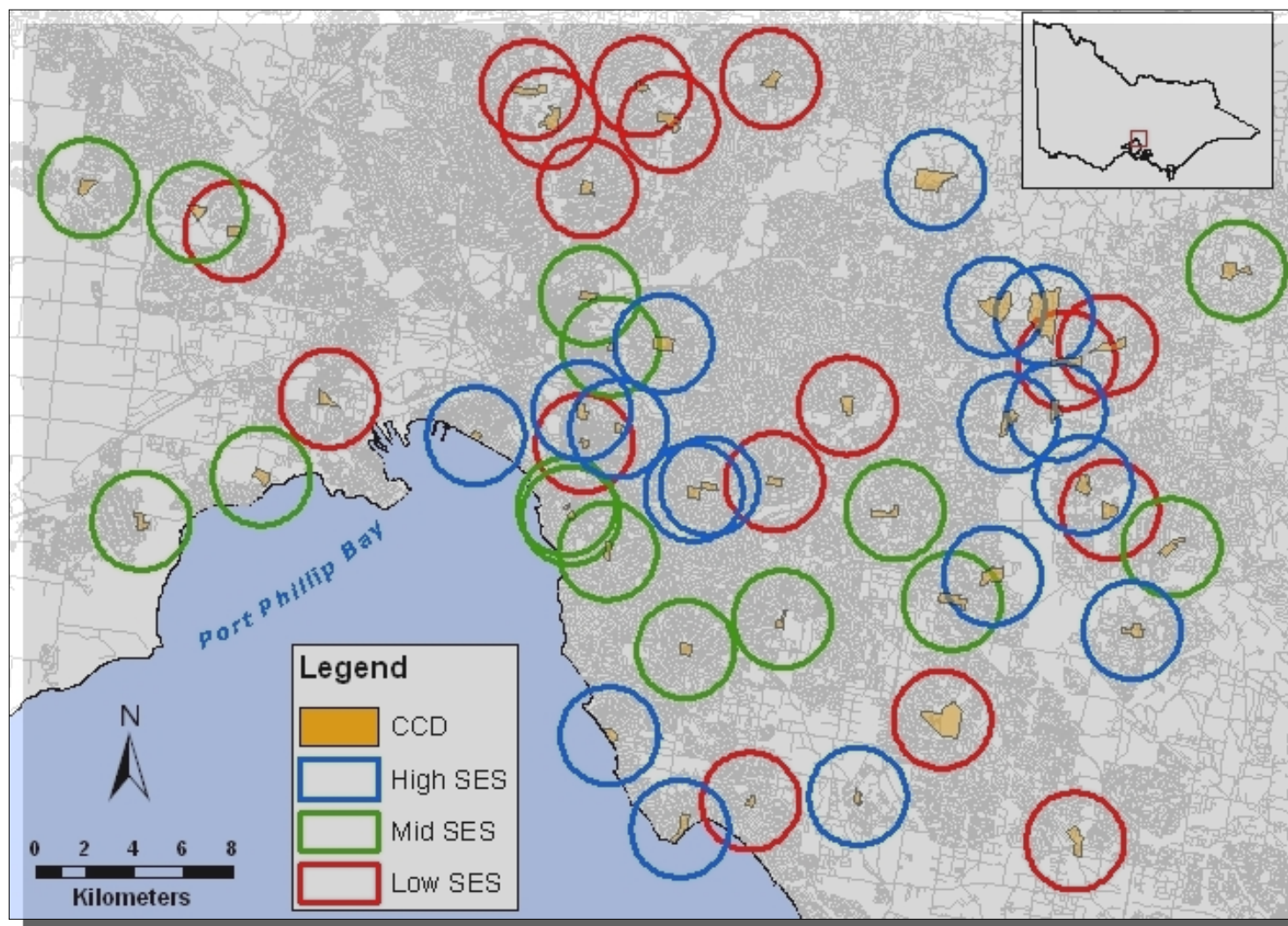


Sample of areas


- High, middle and low SES areas (census collector districts)

Areas	No. CCDs	% low income (<\$400/wk)
High	17	7%
Middle	16	15%
Low	17	31%

VicLANES







Questionnaires



**You and Your
Neighbourhood:**
a Survey




If you have any questions, or require assistance when completing this survey,
please call VicLANES on 1800 008 245



**Food Shopping and
Your Household:**
a survey

If you have any questions, or require assistance when completing this survey,
please call VicLANES on 1800 008 245
www.latrobe.edu.au/viclanes

Sample of individuals

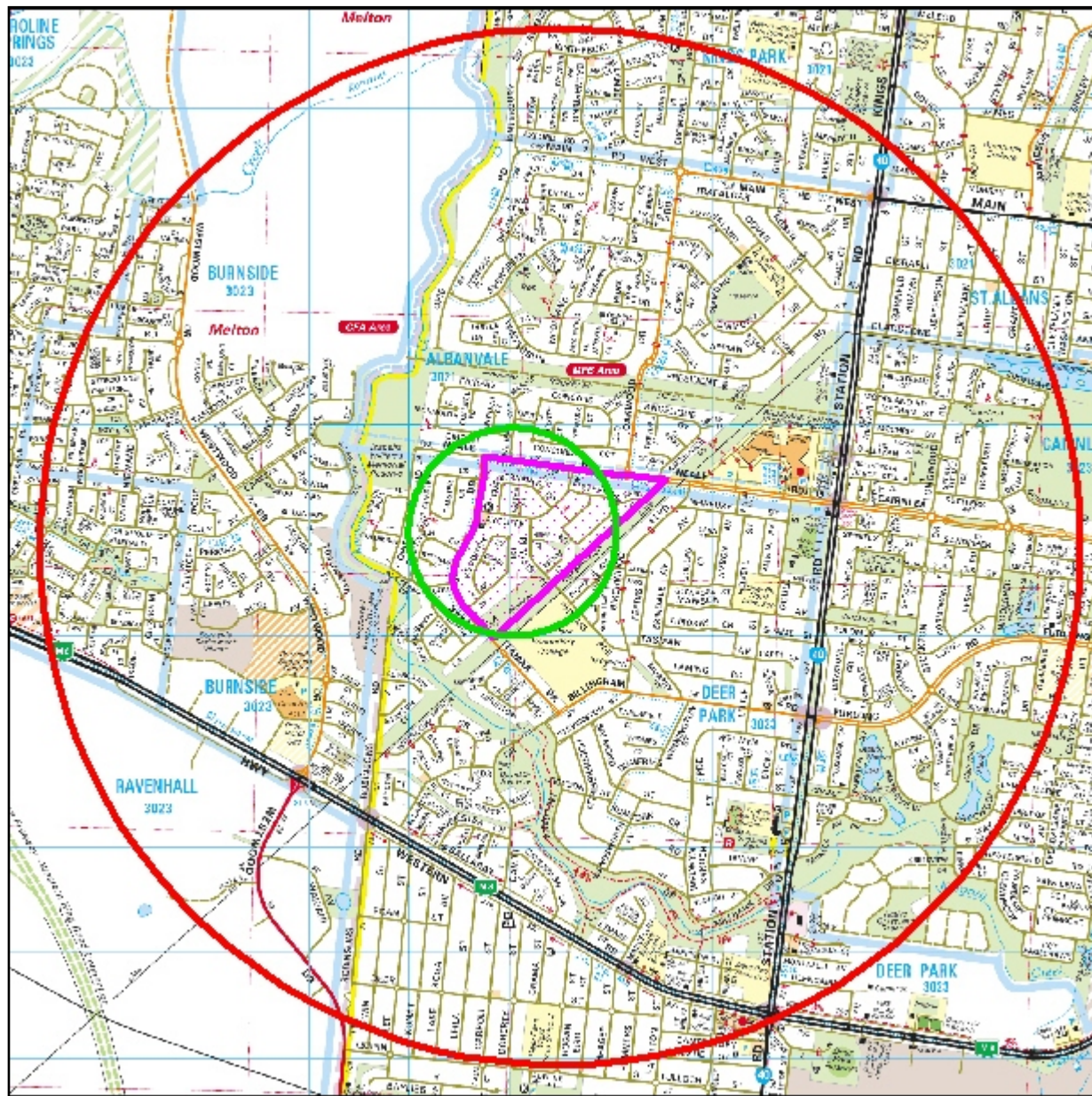
- 4913 individuals: 2564 FP, 2349 PA
- Approx 60% response rate
- 85% food purchasing women

Individual data

- *Outcomes*: physical activity, food-purchasing, alcohol consumption, body mass index
- *Attitudes and knowledge*
- *Perceptions of environment*

Measures of socio-economic status (SES)

- Individual:
 - Household income
 - Education
 - Occupation
- Area SES (% low income)



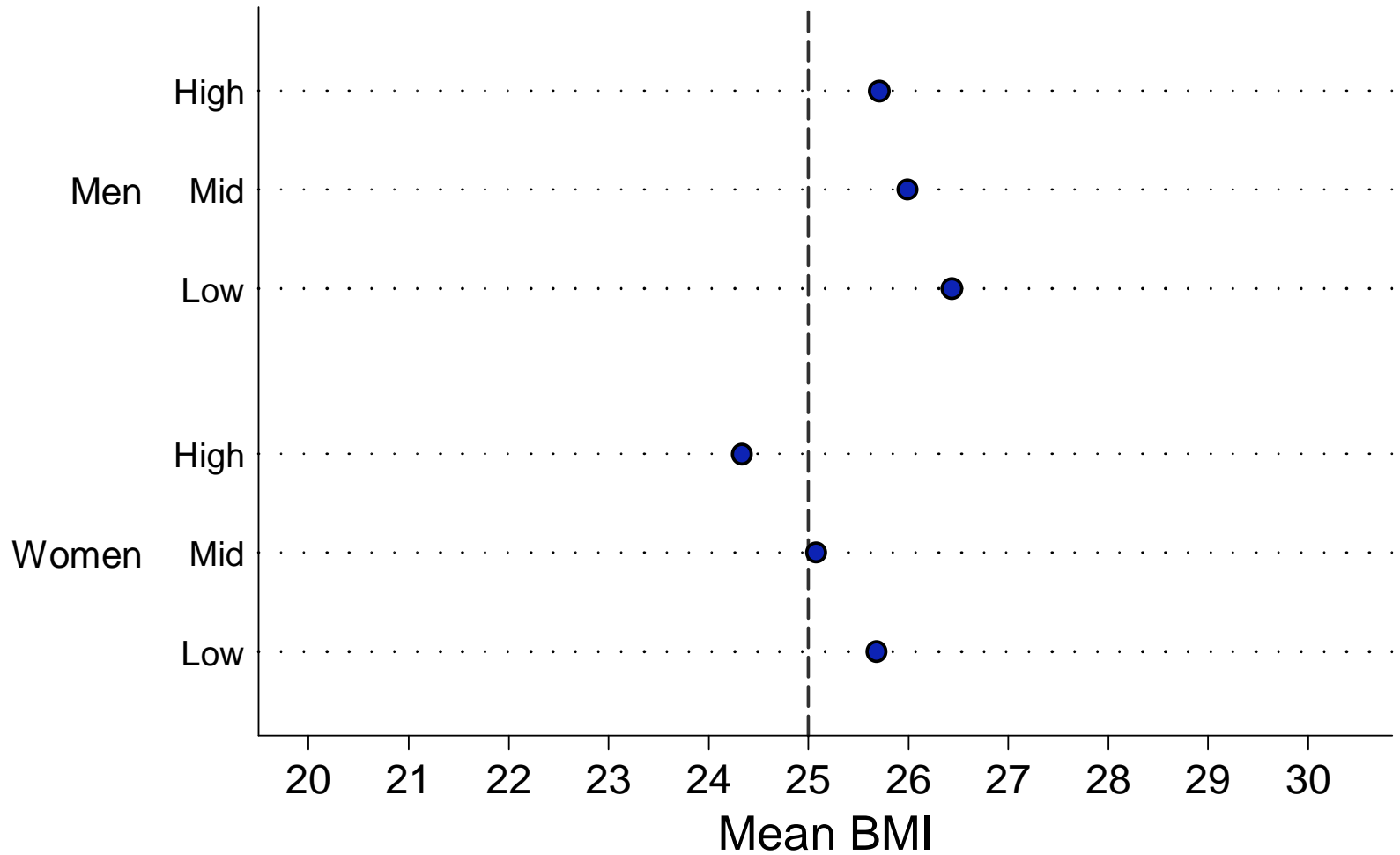
Environmental data: Food

- Identified all stores selling food for consumption in the home in 2km radius
- Collected information on price and availability of 79 different food items in supermarkets, convenience stores, green grocers and ethnic food stores
- 1004 stores audited

Environmental data: physical activity

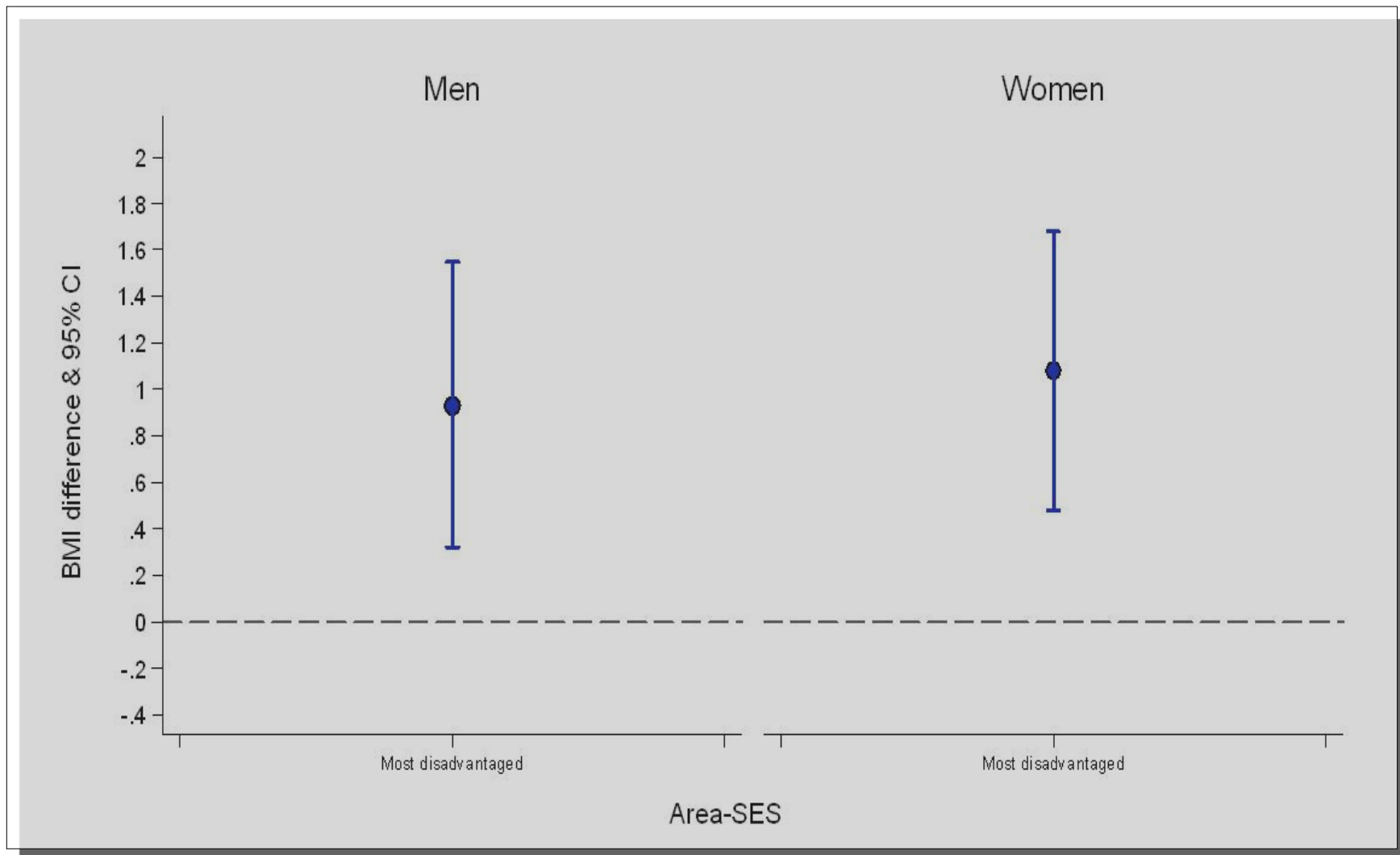
- Structured and unstructured recreational facilities
- Walking and cycling:
 - length of walking and cycling paths
 - audited characteristics in terms of aesthetics, functionality, safety and presence of destinations

BMI



NB: A person with a BMI over 25 is considered overweight or obese

BMI

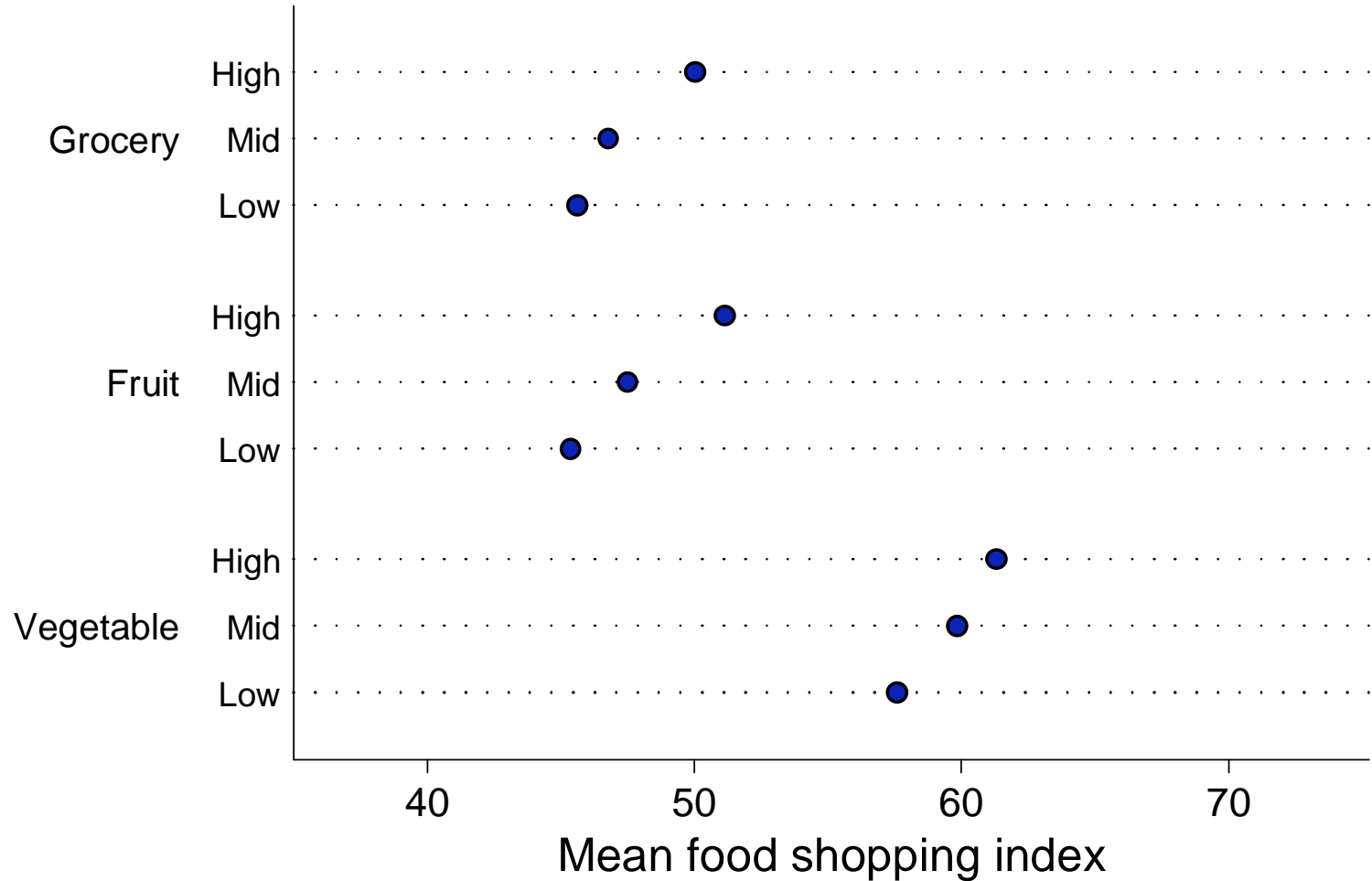


BMI - Random effects

(adjusted for individual SEP & area-SES)

Random effects	Men		Women	
Level 1 (individuals) variance (SE)	14.54	(0.57)	21.62	(0.55)
Level 2 (areas) variance (SE)	0.15	(0.14)	0.36	(0.14)
p-value (Level 2 variance)	0.27		0.01	
ICC	1.05%		1.65%	

Food Purchasing



Grocery index - Random effects

(adjusted for individual SEP & area-SES)

Random effects

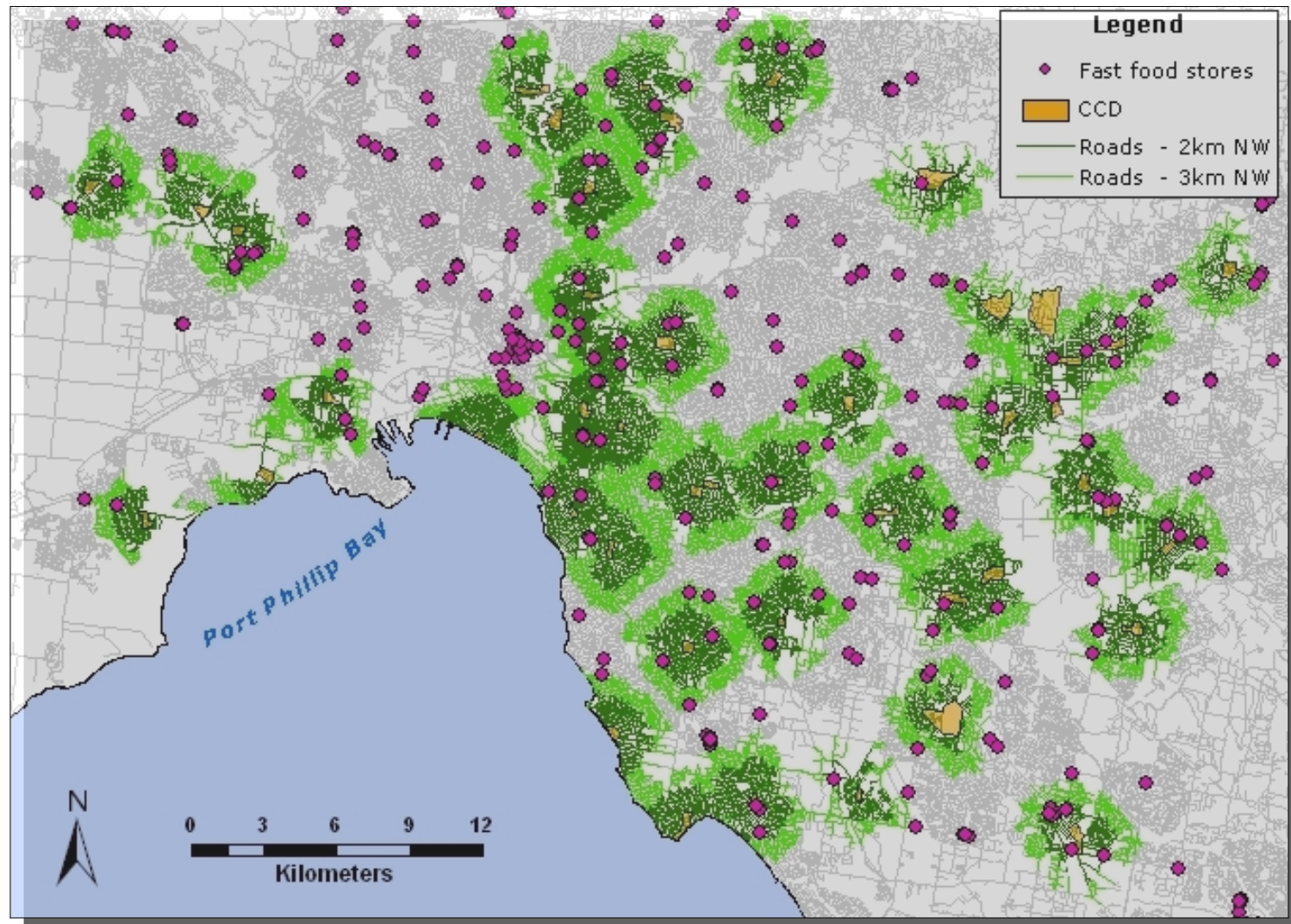
Level 1 (individuals) variance (SE)	166.19	(4.77)
-------------------------------------	--------	--------

Level 2 (areas) variance (SE)	0.42	(0.70)
-------------------------------	------	--------

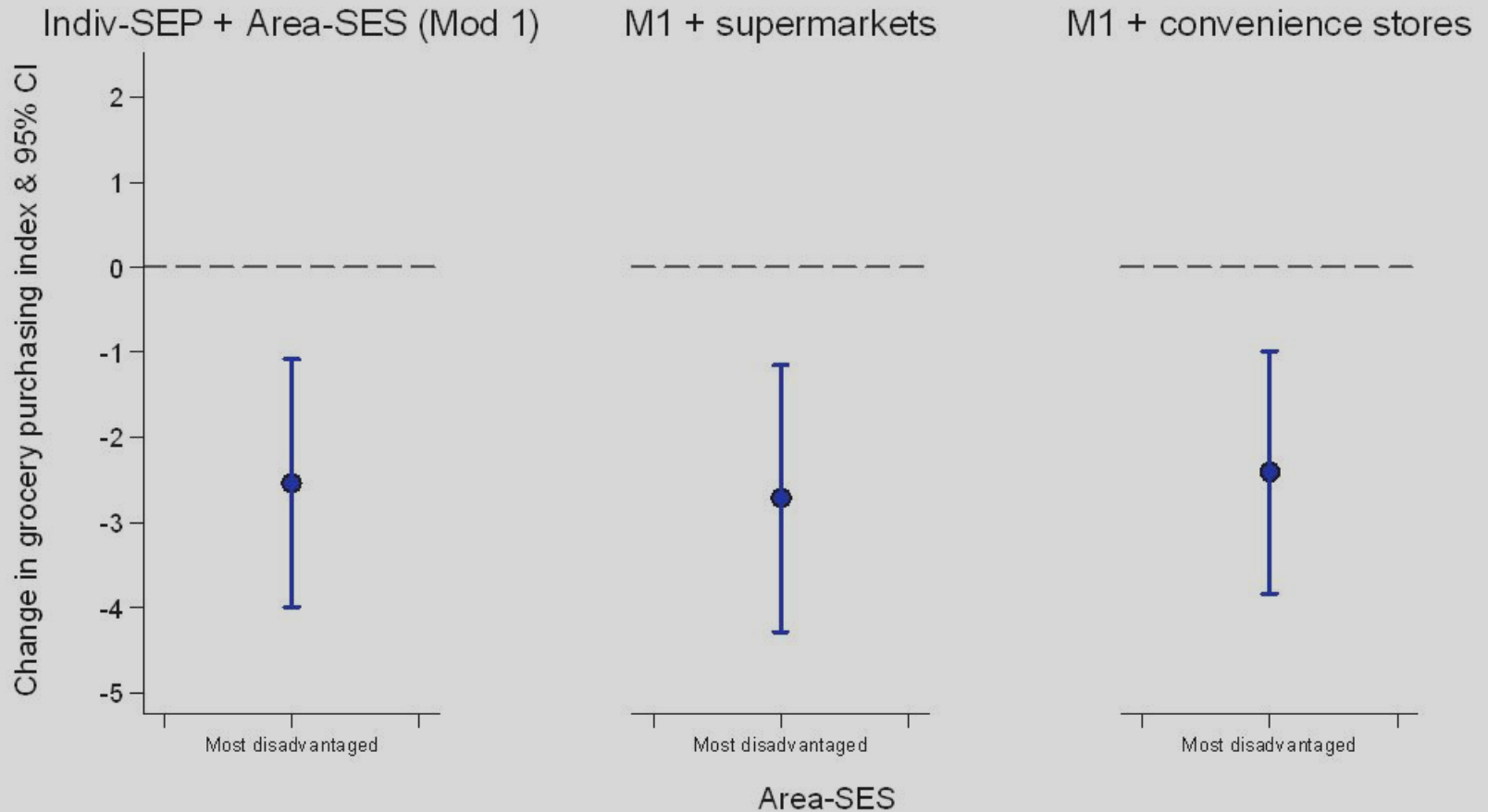
p-value (Level 2 variance)	0.55
----------------------------	------

ICC	0.25%
-----	-------

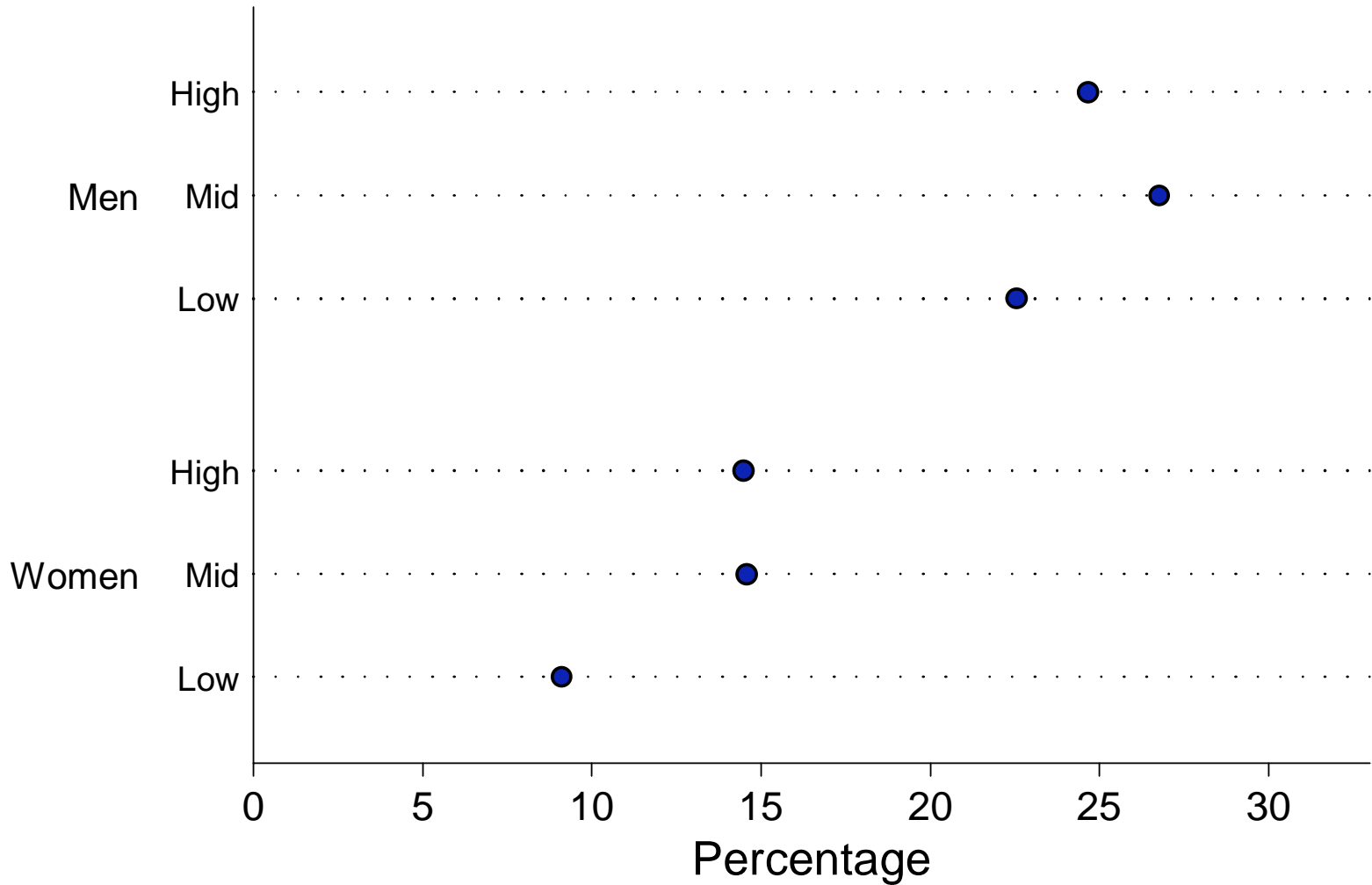
GIS Methods



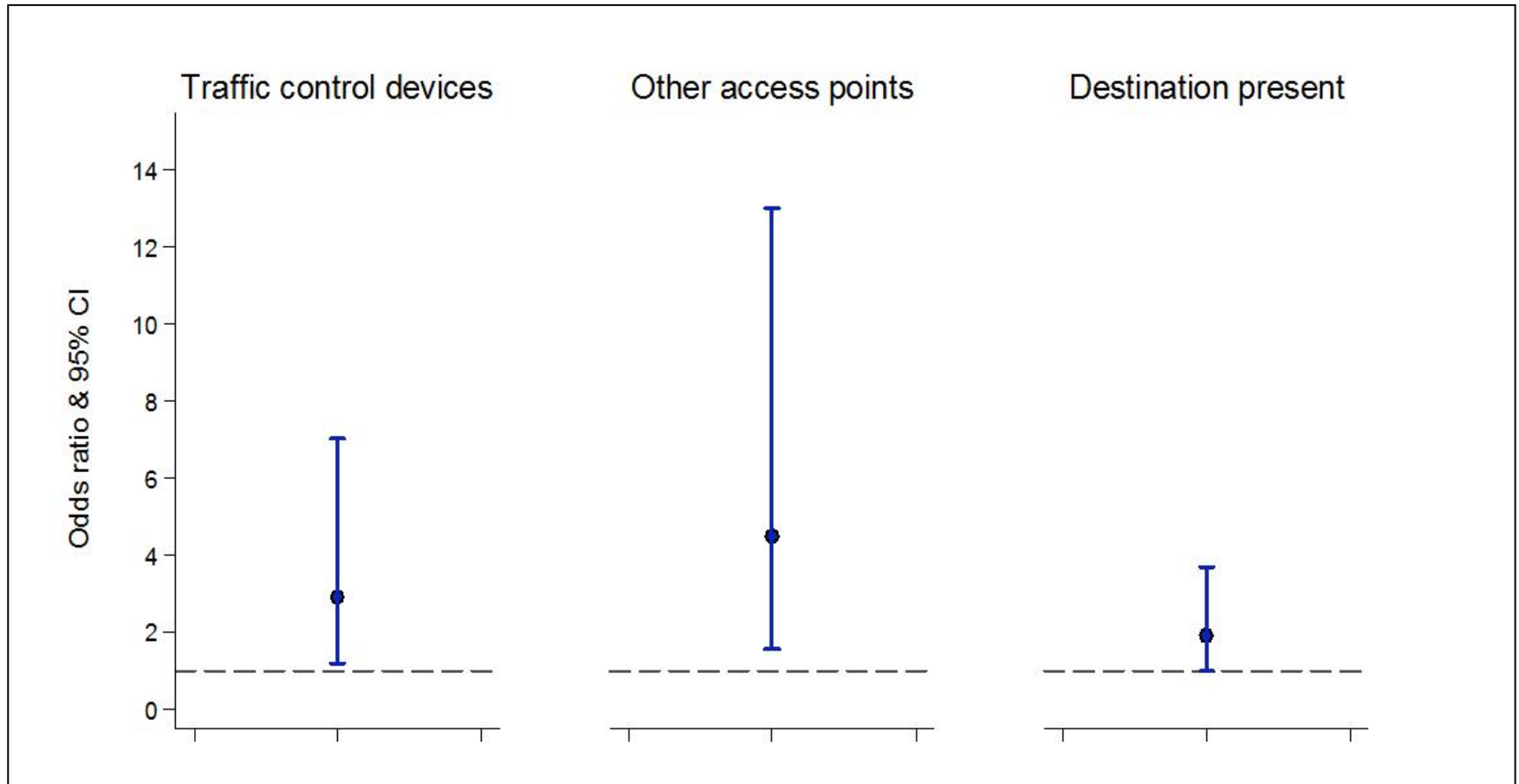
Grocery Index



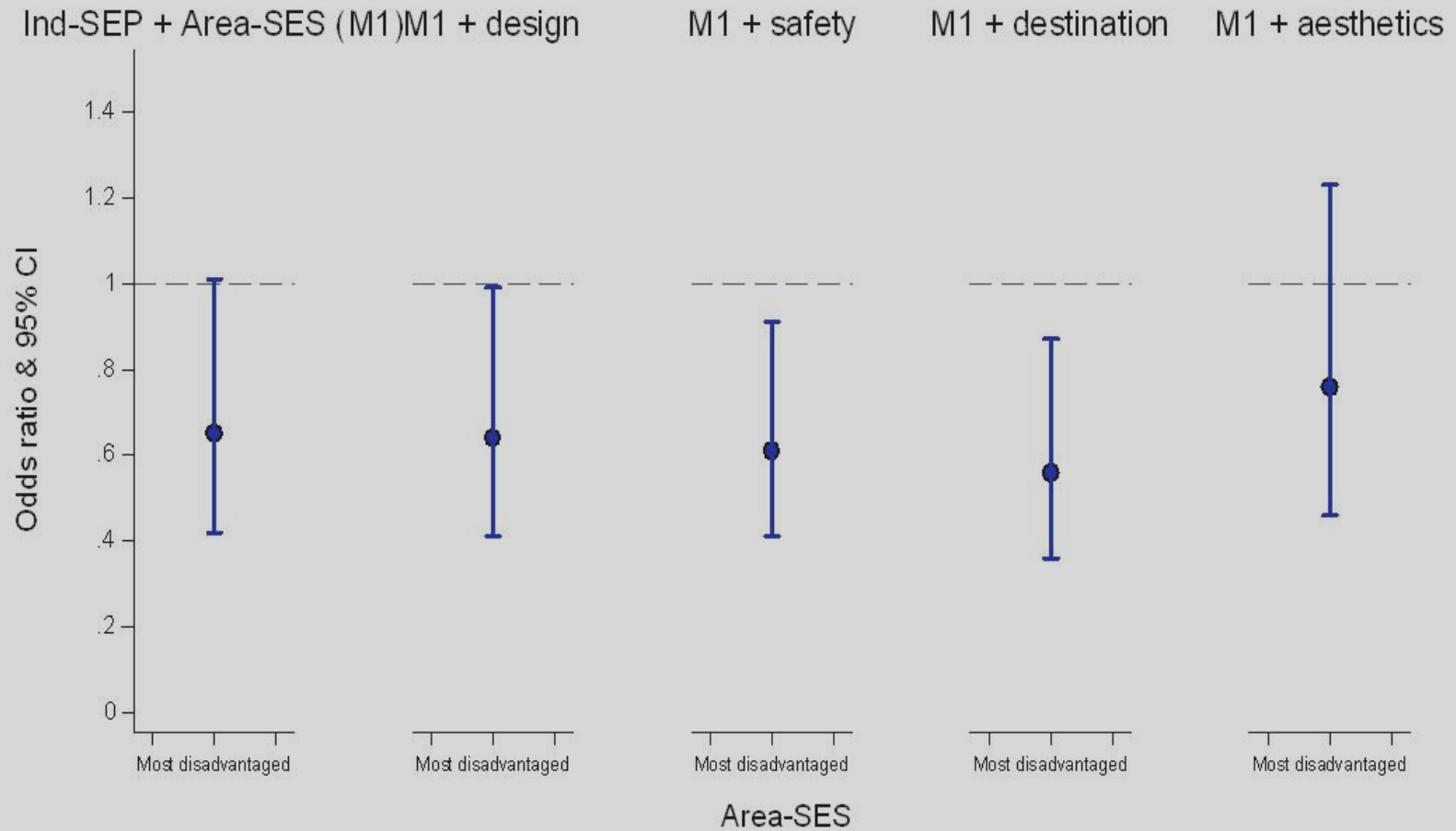
Cycling



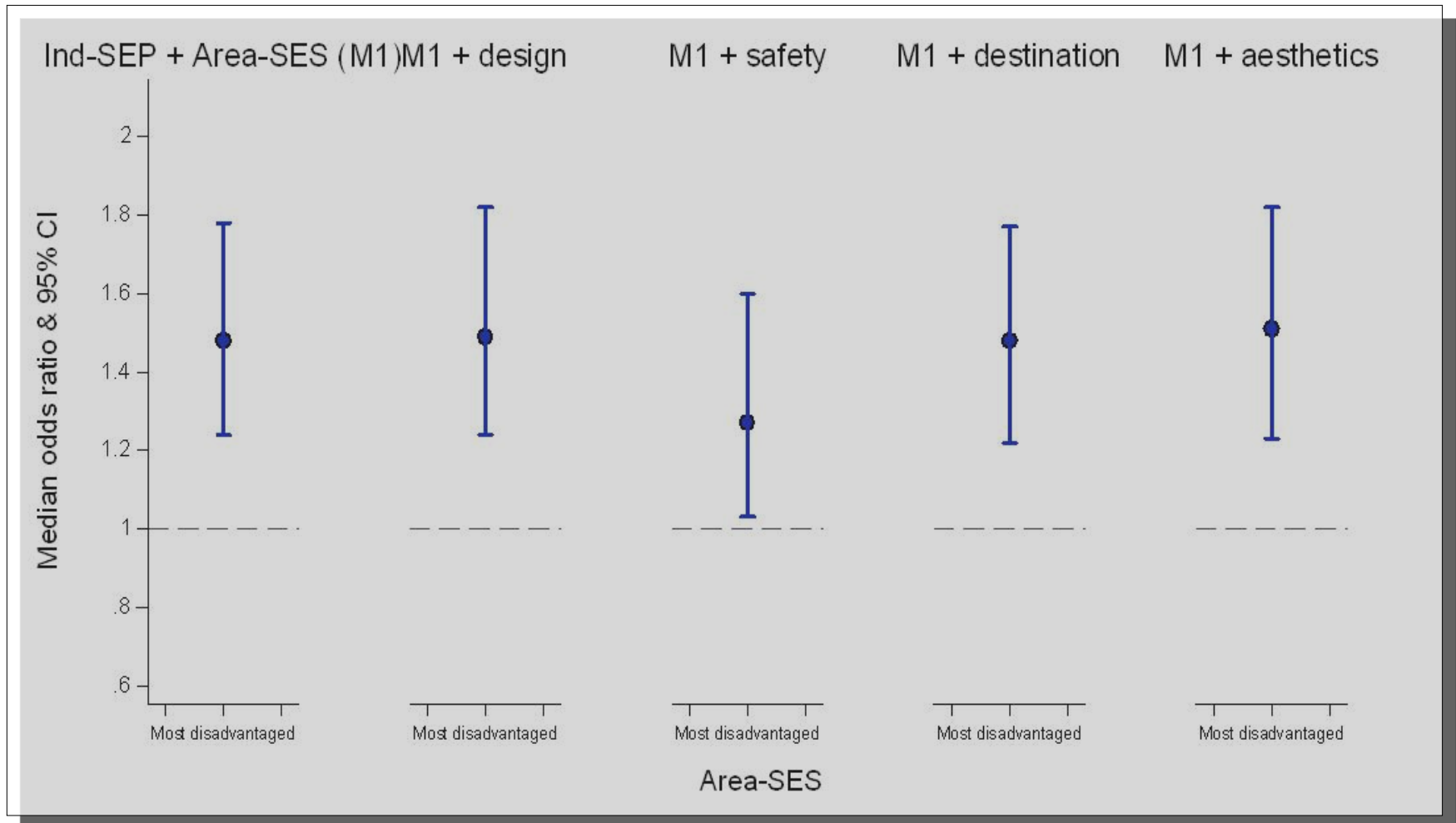
Area characteristics & cycling



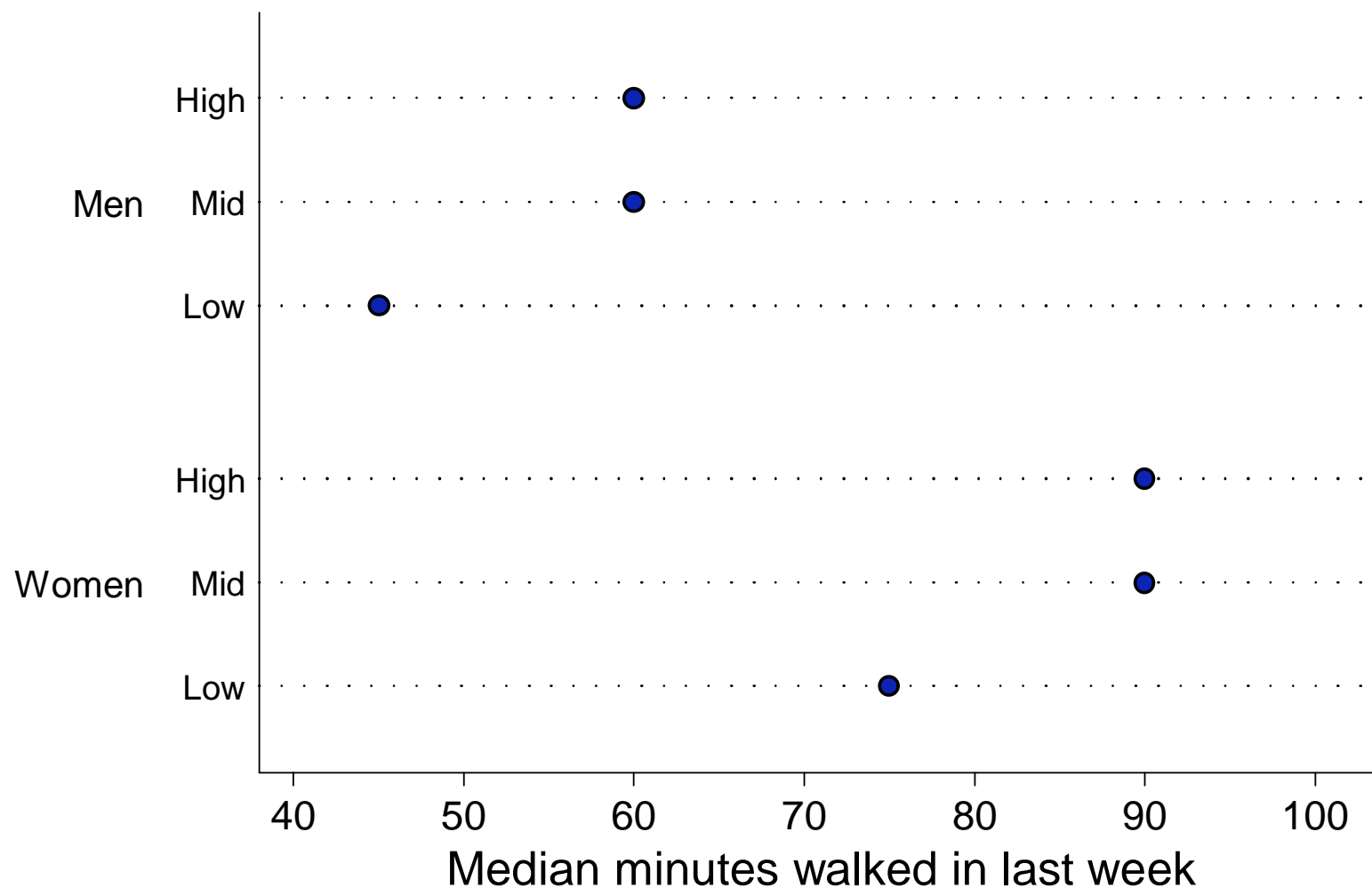
Recreational cycling



Cycling – Median odds ratio

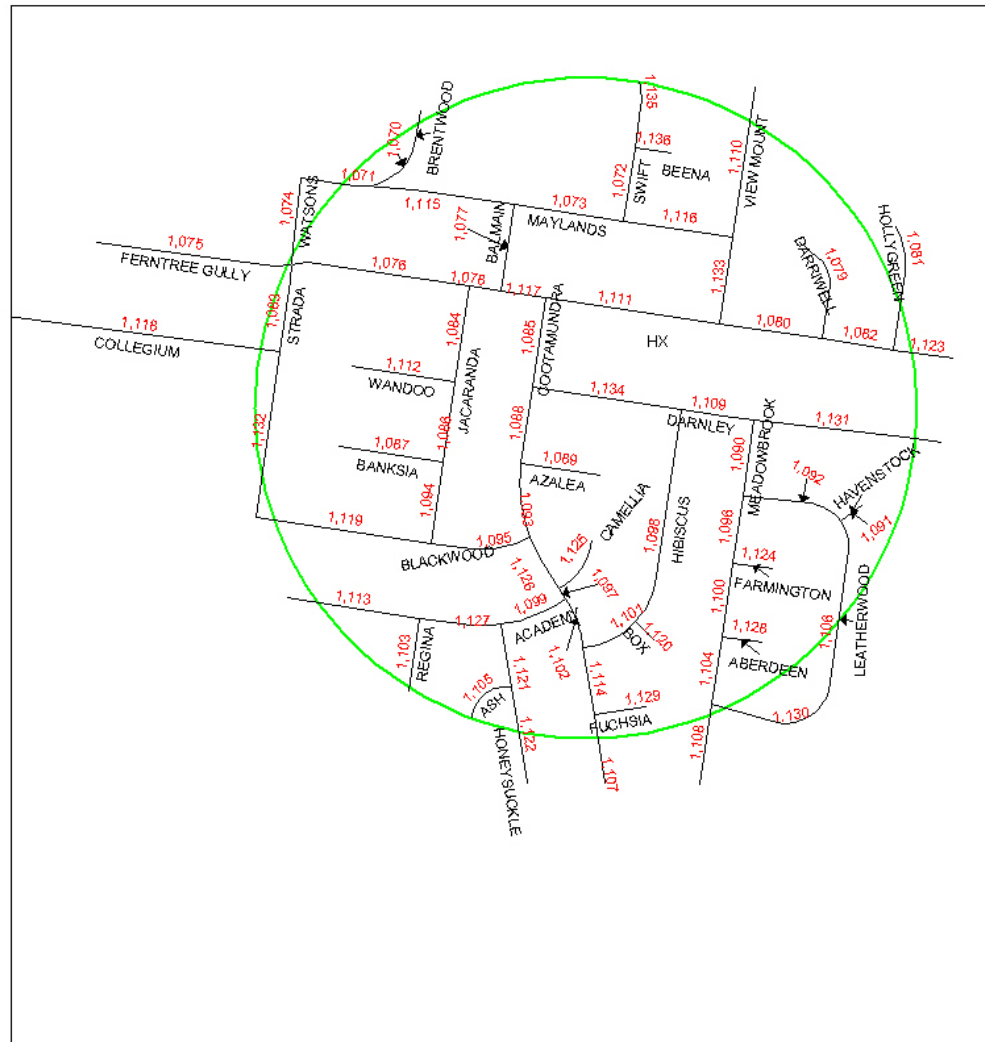


Walking



Factors that might affect walkability

- **Functionality**
 - eg Path type, length, parking restrictions
- **Safety**
 - eg traffic control devices, crossings
- **Aesthetics**
 - eg garden maintenance, types of views
- **Destinations**
 - eg schools, shops, parks, transport, entertainment



- Buffer Zone
- Street
- Street Name
- 22 Segment Number

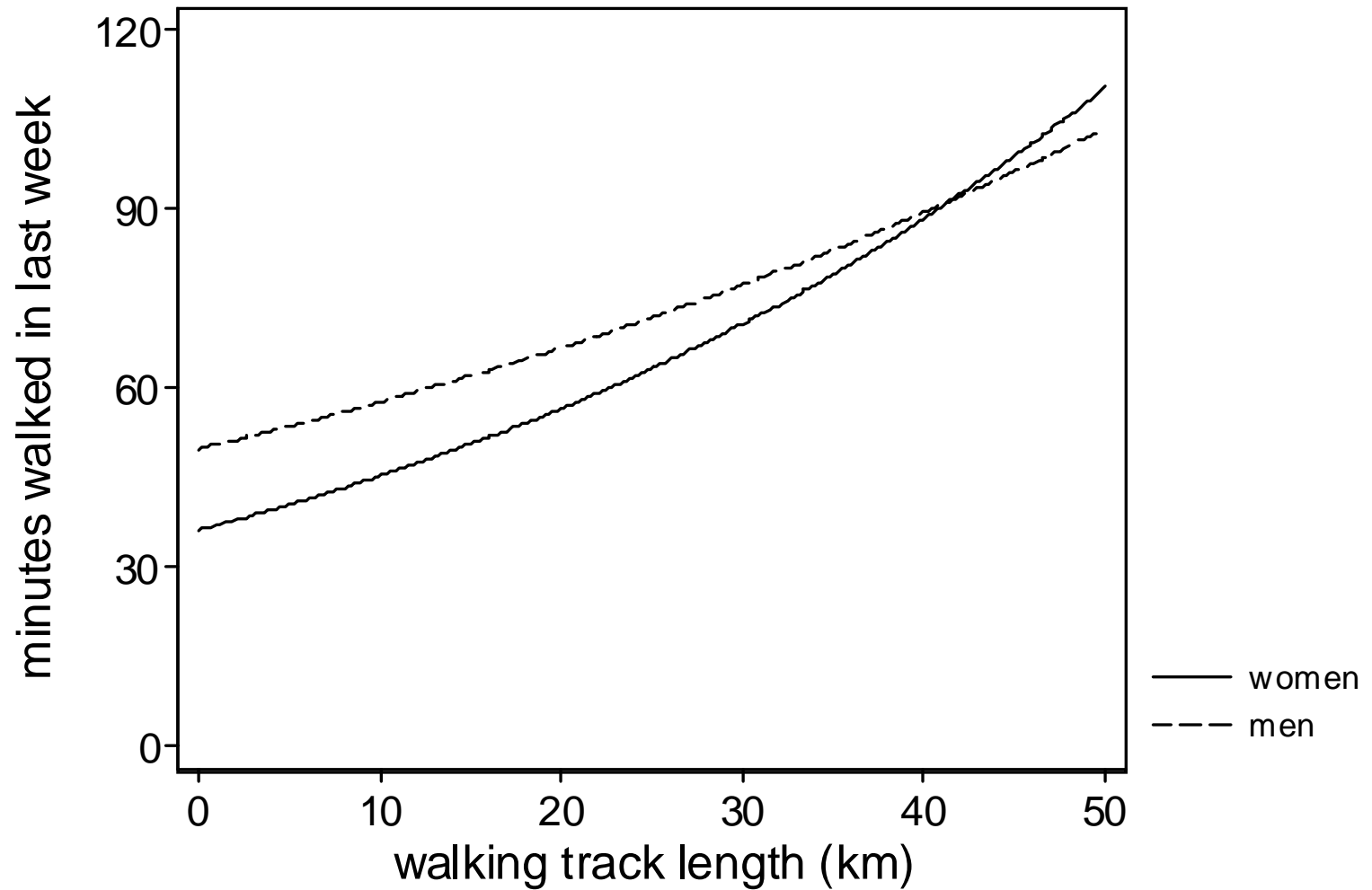
Radius: 400 meters
Scale: 1 cm = 60 meters

Based on Census Collection Districts
and VicMap Data.

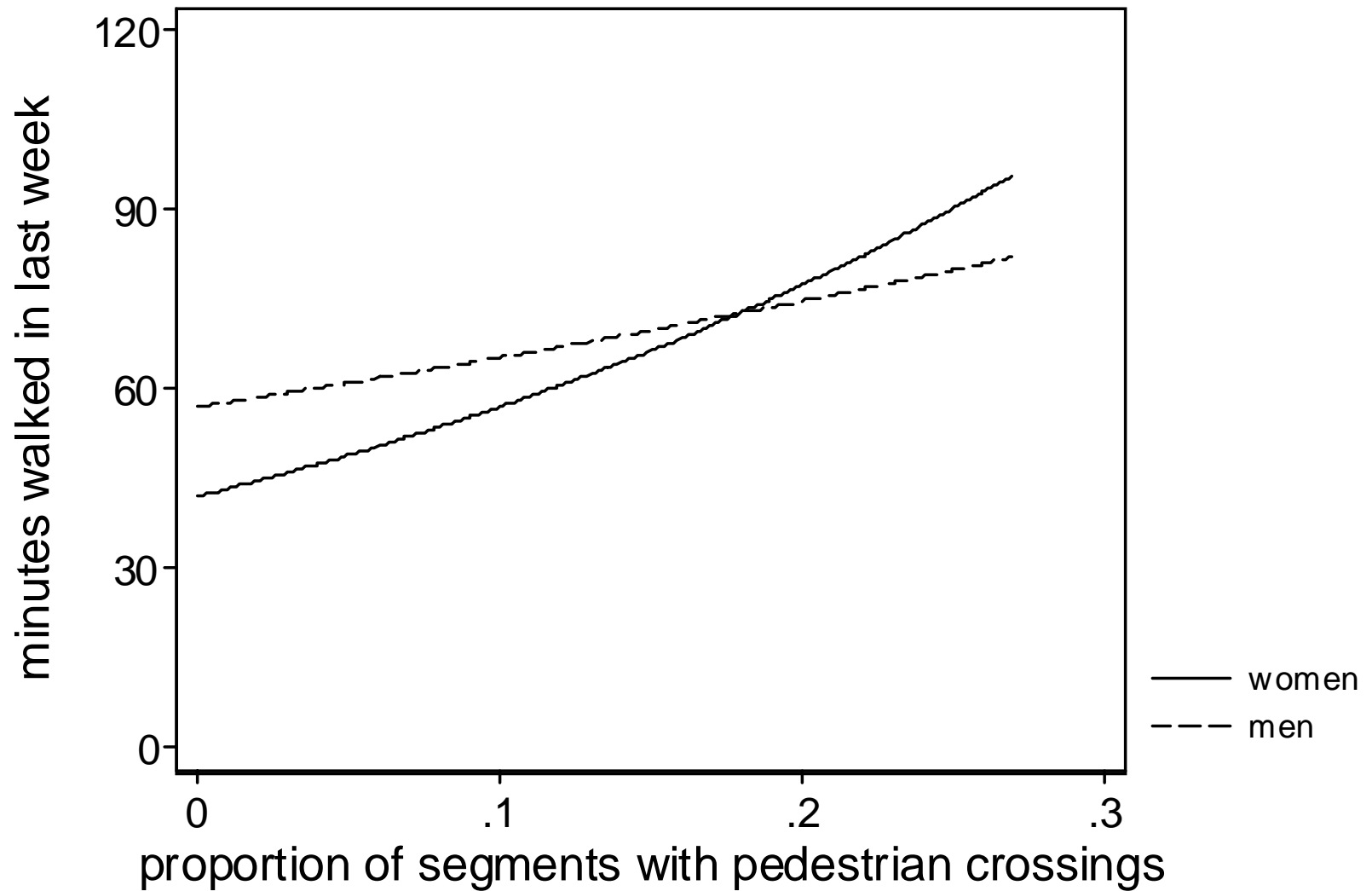
Data from CDATE 2001;
VicMap VMADD_ADDRESS.TAB;
VicMap TR_ROAD.TAB

Produced by A Lovell.
Produced for ARCSHS

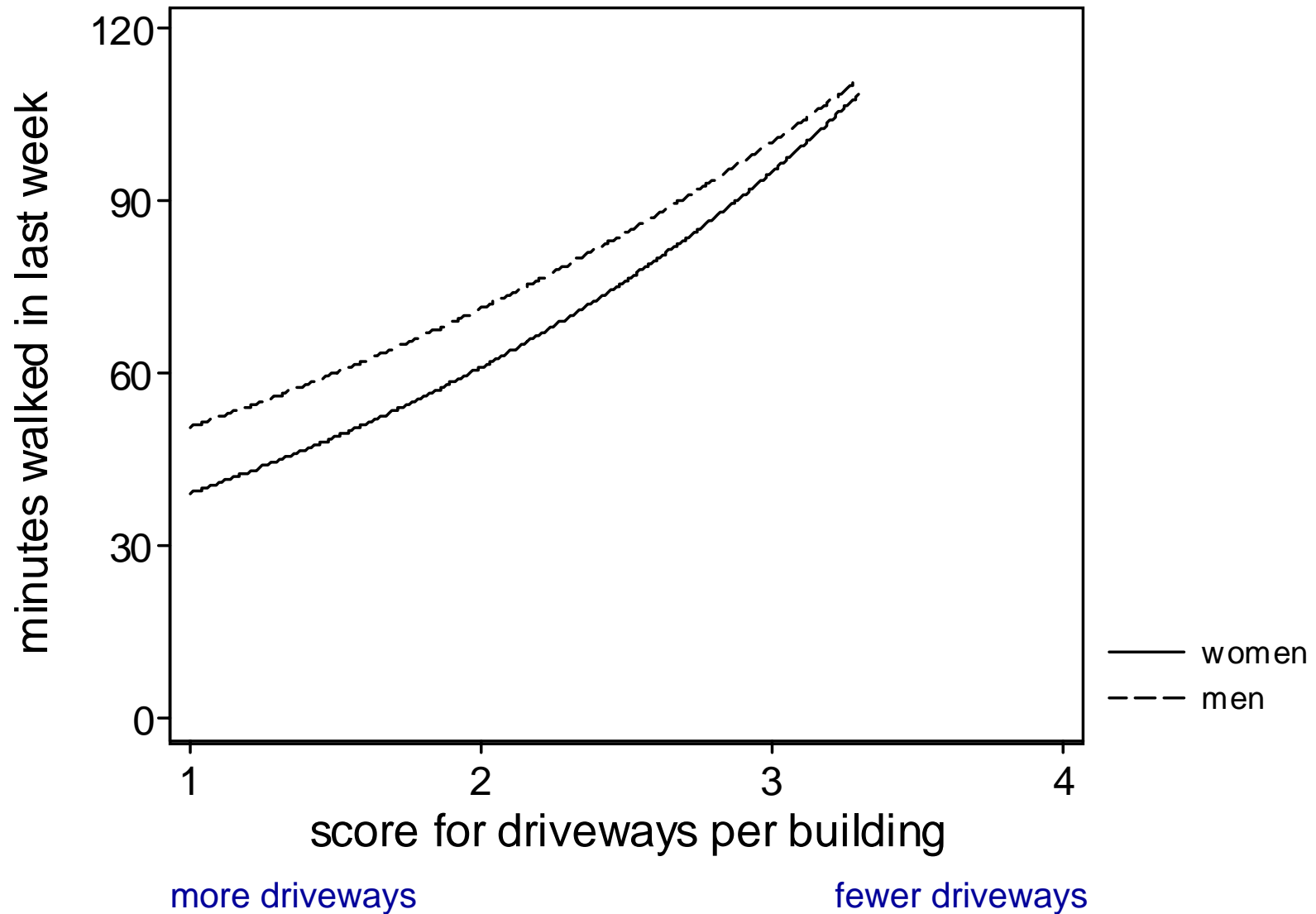
Functional – length of walking tracks



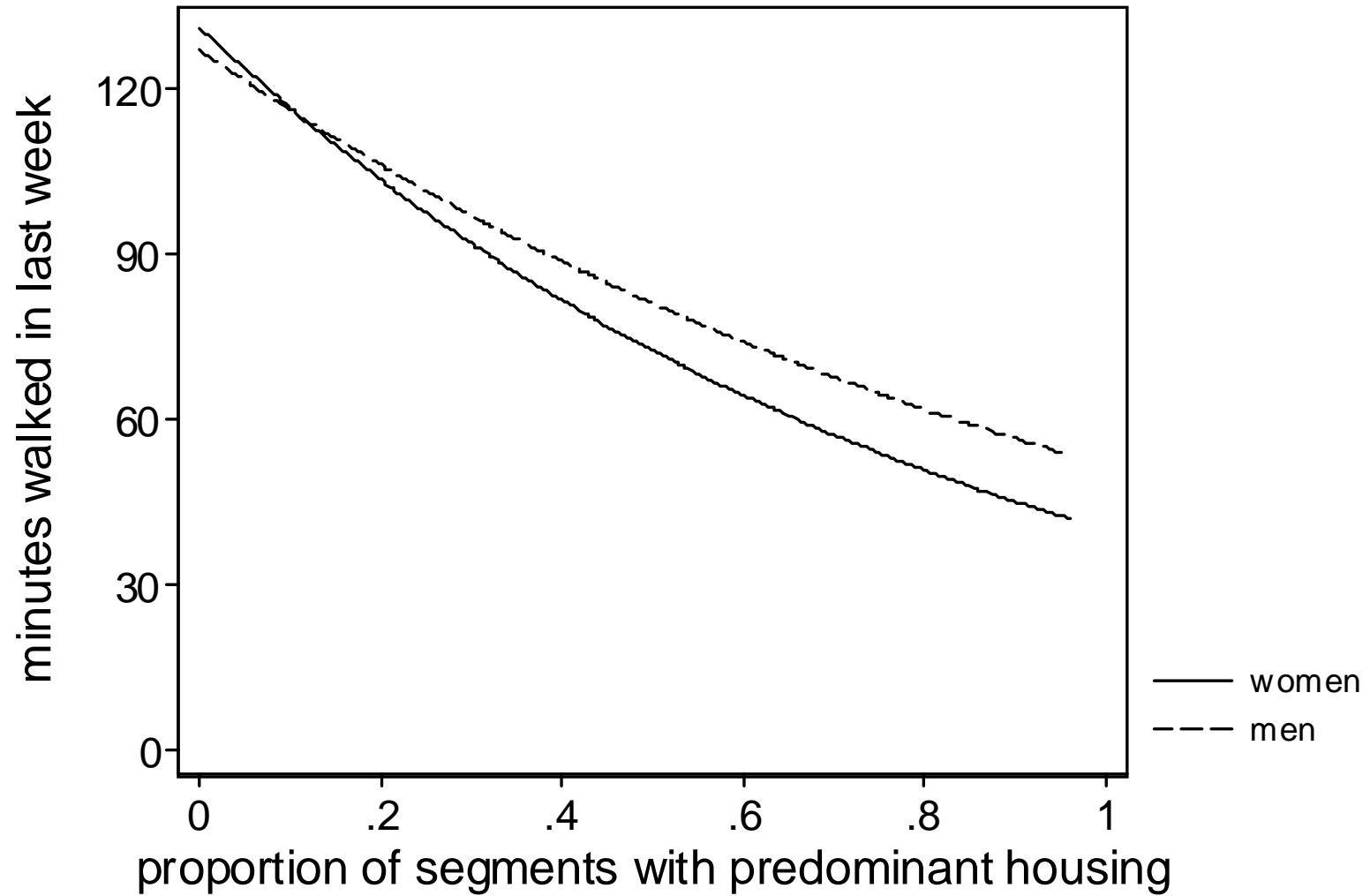
Safety – no. of crossings



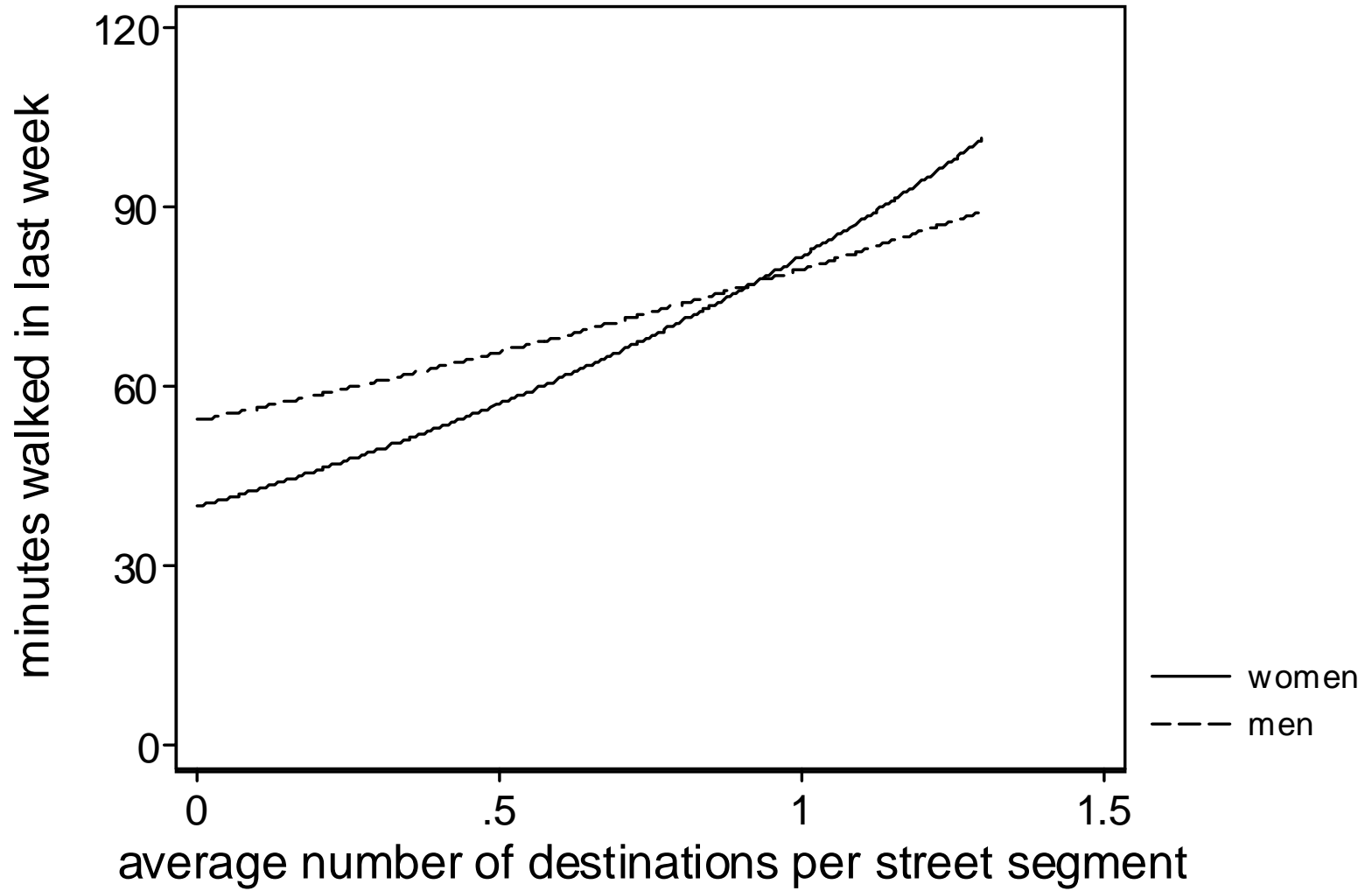
Safety – no. of driveways



Predominant housing



Destinations



Some conclusions

- Between area variation low but some area-level fixed effects
- Some problems with VicLANES:
 - Meaningful area unit specification
 - Insufficient variation in area level exposures
 - Measurement error in area level exposures
 - Too few areas.....

Multilevel models for family and twin studies

Katrina Scurrah

Centre for Molecular, Environmental,
Genetic and Analytic (MEGA)
Epidemiology and
Department of Physiology
University of Melbourne



Family Data

- Genetic association studies or epidemiological family studies
- Outcome + covariates
- Two possible aims:
 - Epidemiological analysis, appropriately adjusting for within-family correlation
 - Use within-family covariance to dissect a “complex” measured outcome (phenotype)
- Complex correlation structure, due to
 - shared genes
 - shared environmental influences

THREE PHASES OF GENETIC EPIDEMIOLOGY

Understanding

Aim:

Is there a genetic aetiology?
How strong is it?
Mode of inheritance, etc.

Genes:

Unmeasured

Optimal Design:

Population-based

Families, twins, adoptees

Statistical Analysis:

Pedigree

incl. segregation analysis

Discovery

Where are the genes?
What are the variants?

Inferred from
markers

Opportunistic

Families, affected pairs

Linkage

model known/unknown

Characterisation

What do they mean?
Penetrance (risk)
Prevalence (how much)

Measured
incl. candidates

Population-based

Individuals & families

Epidemiology

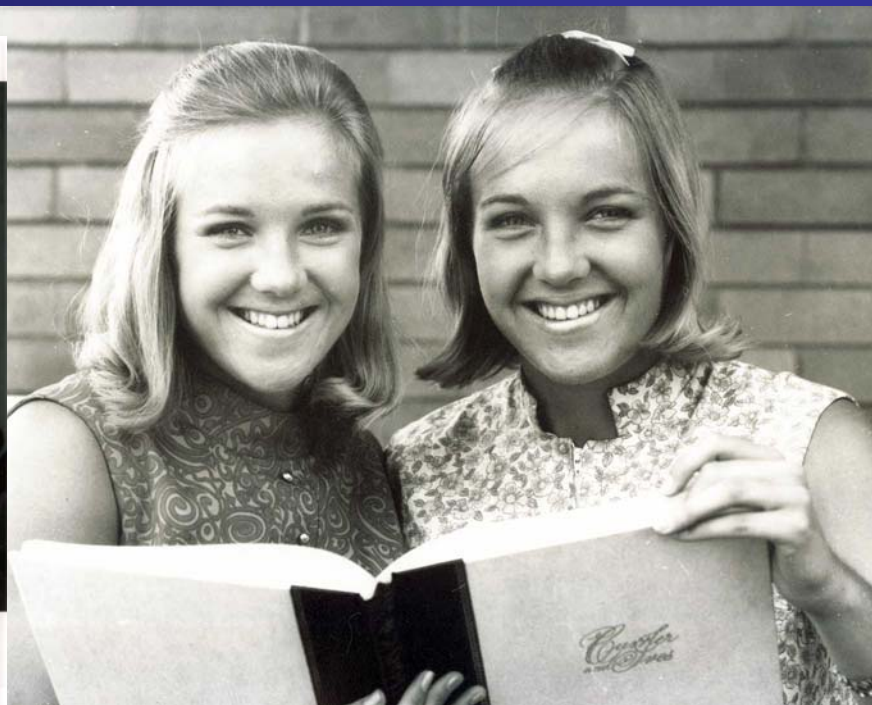
incl. family analysis

Shared genes

- Each parent passes down exactly half their genes to each child
- Each pair of siblings and dizygous (DZ, non-identical) twins shares half their genes (on average)
- Monozygous (MZ, identical) twins share all their genes

Shared environments

- Families tend to have similar diets, exercise habits etc
- Sibling pairs may be more similar than parent-offspring pairs
- Twin pairs may be even more similar!



Variance components models

- Type of multilevel model widely used in genetic epidemiology
- Multivariate normal distribution assumed
- Fitted using maximum likelihood estimation
- Correlation models simply estimate correlation for each relative pair type
- Variance components models partition residual phenotypic variance into components due to
 - Additive polygenic effects
 - Shared family environmental effects
 - Unshared environmental effects
- Provide information about contributions of shared genes and shared environment
- Account for covariates and family structures
- Genotypes not required

Simple variance components model

For each family i ,

$$\underline{y}_i \sim N(\beta^T X_i, V_i)$$

$$V_i = \sigma^2 \times R_i$$

σ^2 = variance of y

	R_i				
	F	M	S1	MZ1	MZ2
F	1	ρ_{SP}	ρ_{PO}	ρ_{PO}	ρ_{PO}
M		1	ρ_{PO}	ρ_{PO}	ρ_{PO}
S1			1	ρ_{SIB}	ρ_{SIB}
MZ1				1	ρ_{MZ}
MZ2					1

Full variance components model

$$\underline{y}_i \sim N(\beta^T X_i, V_i)$$

$$V_{i,jk} = \begin{cases} 2\phi_{jk}\sigma_A^2 + \gamma_{jk}\sigma_C^2, & j \neq k \\ \sigma_A^2 + \sigma_C^2 + \sigma_E^2, & j = k \end{cases}$$

ϕ_{jk} =kinship coefficient

(0.5 for MZ twins, 0.25 for first degree relatives)

γ_{jk} =environmental coefficient

Variance components:

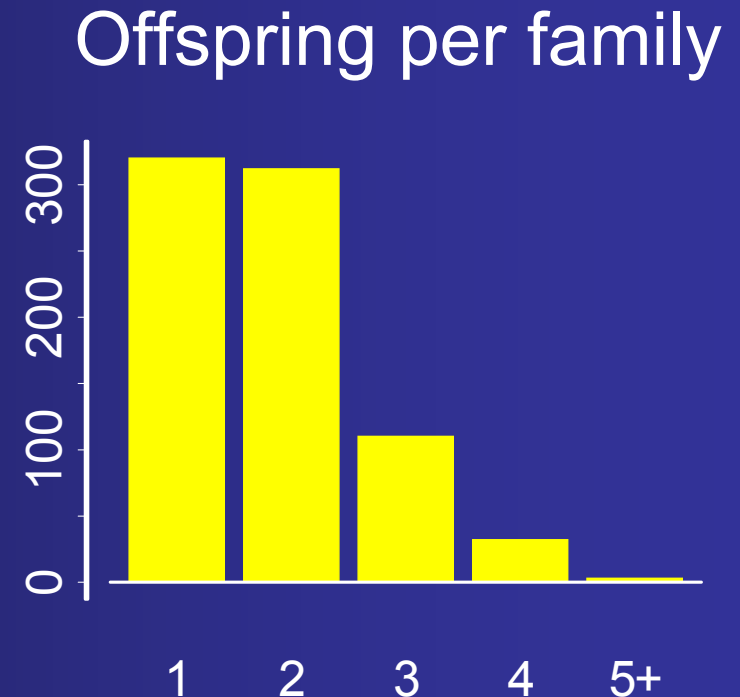
σ_A^2 (additive polygenic), σ_C^2 (common family environmental), σ_E^2 (individual-specific)

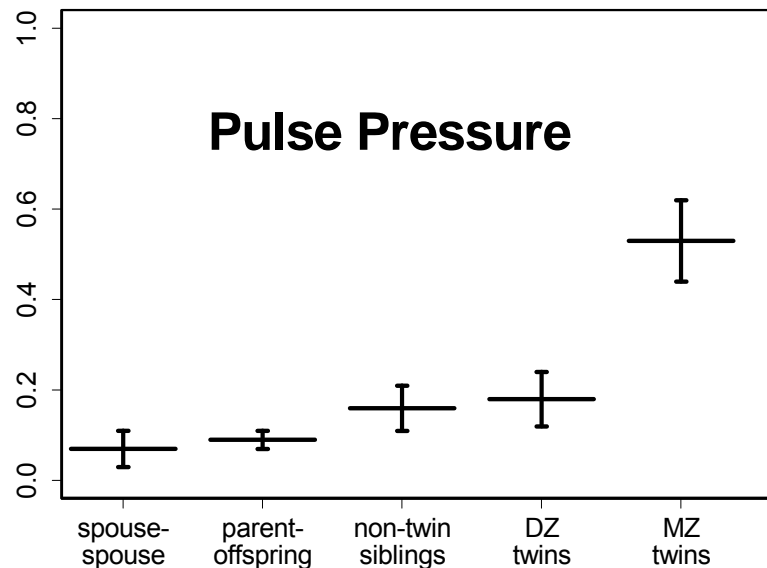
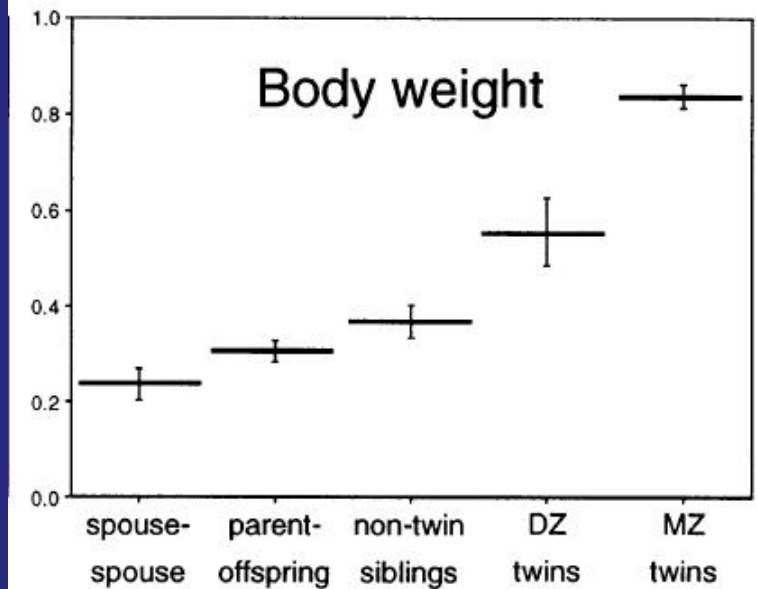
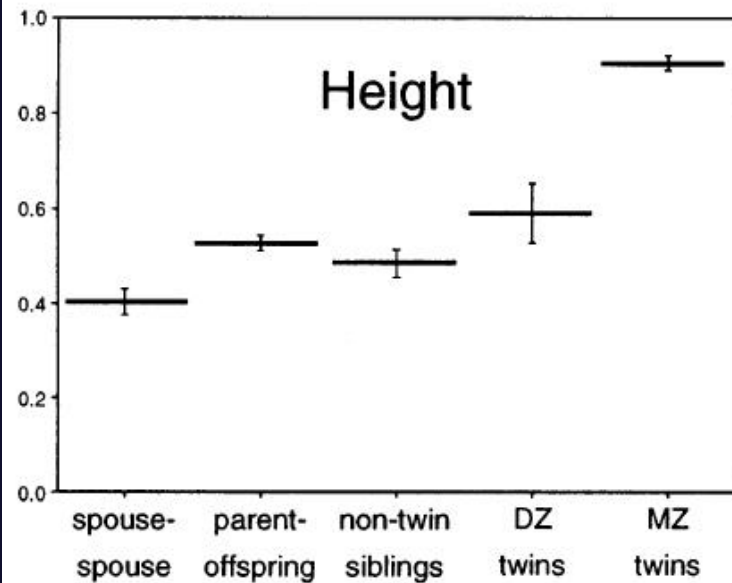
Full variance components model: V_i

	F	M	S1	MZ1	MZ2
F	$\sigma^2_A + \sigma^2_C + \sigma^2_E$	$\rho_{SP} \times (\sigma^2_A + \sigma^2_C + \sigma^2_E)$	$\frac{1}{2}\sigma^2_A + \gamma_{PO} \times \sigma^2_C$	$\frac{1}{2}\sigma^2_A + \gamma_{PO} \times \sigma^2_C$	$\frac{1}{2}\sigma^2_A + \gamma_{PO} \times \sigma^2_C$
M		$\sigma^2_A + \sigma^2_C + \sigma^2_E$	$\frac{1}{2}\sigma^2_A + \gamma_{PO} \times \sigma^2_C$	$\frac{1}{2}\sigma^2_A + \gamma_{PO} \times \sigma^2_C$	$\frac{1}{2}\sigma^2_A + \gamma_{PO} \times \sigma^2_C$
S1			$\sigma^2_A + \sigma^2_C + \sigma^2_E$	$\frac{1}{2}\sigma^2_A + \gamma_{SIB} \times \sigma^2_C$	$\frac{1}{2}\sigma^2_A + \gamma_{SIB} \times \sigma^2_C$
MZ1				$\sigma^2_A + \sigma^2_C + \sigma^2_E$	$\frac{1}{2}\sigma^2_A + \sigma^2_C$
MZ2					$\sigma^2_A + \sigma^2_C + \sigma^2_E$

Victorian Family Heart Study

- Aim: to identify genetic influences on cardiovascular risk factors
- Nuclear families enriched with twin families
- Traits for 2911 individuals in 767 nuclear families
 - height,
 - weight,
 - PP (=SBP-DBP)

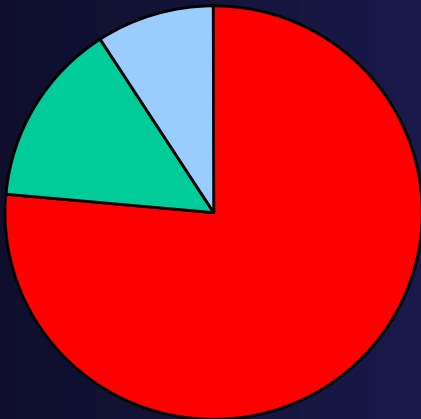




Figures for height and weight from Harrap et al., 2000, American Journal of Epidemiology

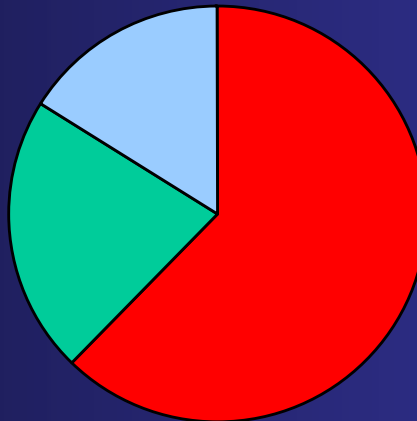
Variance Components Analyses

Height



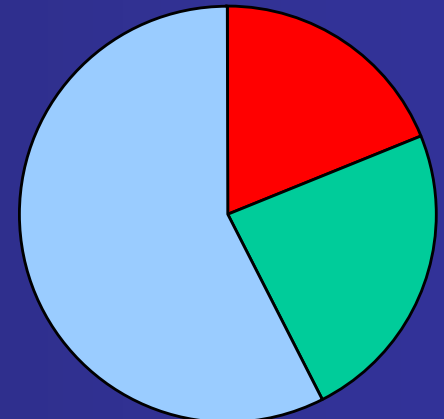
σ^2_A
Shared
polygenic
effects

Weight



σ^2_C
Shared
environmental
effects

Pulse
Pressure



σ^2_E
Unshared
environmental
effects

Additional issues

- Extra covariate effects
 - SEP
 - Measured genotype data
 - Genome wide linkage scans, SNP association studies, haplotype analyses
- Assortative mating for height
- Non-normal outcomes
 - Male pattern baldness, prostate cancer, age of onset of these
 - Use GLMMs fitted using WinBUGS
 - Binary outcomes (Burton et al. 1999), survival times (Scurrah et al. 2000), ordered categorical traits (current work)

Hints and tips

- Start with simple models and add complexity
 - linear regression → single family random effect → correlation model → full VC model
 - Use estimates from simple models as starting values for more complex models
- Complexity of data determines complexity of models that can be fitted
- Specialised software may be required due to complex genetic sharing
 - Fisher, Solar, “kinship” package in R
- Aim for robustness and consistency
- Check for outliers at multiple levels

Summary

- Multilevel models are extremely useful when analysing family data, whether focus is on fixed or random effects, and in all phases of genetic epidemiology

Study Investigators and Acknowledgements

Harrap Lab

Stephen Harrap
Justine Ellis
Zilla Wong
Sophie Zaloumis
Joanna Cobb
Cara Büssst
Anna Duncan
Tania Infantino
Angela Lamantia
Leona Yip

MEGA Epidemiology Centre

John Hopper
Graham Byrnes
Lyle Gurrin

Melanie Bahlo
Graham Giles
Hoa Hoang
Rod Sinclair
James Cui
Margaret Stebbing

Health 2000

CSIRO

GPs

NHMRC

NHF

Victorian Health Promotion Foundation

Australian Twin Registry

Study Participants

Research Nurses

Cluster randomised trials

Obi Ukoumunne

CEBU

Murdoch Childrens Research Institute

Overview

- cluster randomised trials (CRTs)
- general analytical issues in CRTs
- application of multilevel models to the “Gatehouse” trial
- cluster-specific (CS) versus population-averaged (PA) estimates of intervention effect
- other “clustered” designs

Cluster randomised trials (CRTs)

- In traditional randomised controlled trials individuals are randomised to trial arms
- In CRTs
 - clusters (groups of subjects) are randomised
 - outcomes measured on individual subjects
 - subjects in the same cluster are effectively randomised to the same trial arm

Cluster randomised trials (CRTs)

- Clusters may be:
 - health organisational units
 - general practices, hospitals
 - health professionals
 - doctors, maternal and child health nurses
 - geographic areas
 - local government areas
 - non-health organisations
 - schools, offices

Example: Gatehouse Study

- **Aim:** Effectiveness of an intervention for increasing social inclusiveness in schools
- **Subjects:** Children aged 13 to 14
- **Intervention:** Delivered in schools
- **Outcomes:** Heavy substance use
- **Randomisation clusters:** 11 intervention schools and 14 control schools
- **Observation units:** 2579 children (average of 103 per cluster)

Analytical issues in CRTs

- subjects in the same cluster are effectively randomised to the same trial arm
- outcomes of subjects from the same cluster tend to be correlated within clusters
- equivalently there is an additional source of outcome variation that is between clusters
- standard statistical methods invalidated
 - they incorrectly assume independence
 - do not recognise between-cluster variation

Analytical issues in CRTs

- standard analytical methods that don't allow for clustering will usually:
 - underestimate the variance of the intervention effect estimate
 - produce confidence intervals that are too narrow and p-values that are too small
 - produce intervention effect estimates that are (unbiased but) not maximally efficient if the number of individuals varies between clusters

What multilevel modelling offers

- nested data structure of CRTs is naturally suited to multilevel analysis
 - level 1 units – individuals/subjects
 - level 2 units – clusters (randomisation units)
- intermediate levels of clustering between level 1 and level 2 may be ignored in specification of multilevel model
 - e.g. if general practices are randomised, doctors (or other health professionals) are a level of clustering nested between the practice and patient levels

Multilevel model to estimate intervention effect

- Multilevel model for comparing two trial arms w.r.t. a continuous outcome in CRT

$$y_{ij} = \beta_0 + \beta_1 * G_i + u_i + e_{ij}$$

- y_{ij} – outcome of j th subject in i th cluster
- G_i – indicator variable for trial arm status
- u_i – random effect of i th cluster $\sim N(0, \sigma^2_u)$
- e_{ij} – residual effect $\sim N(0, \sigma^2_e)$
- β_1 – mean difference between trial arms

Example: Gatehouse Study

- **Aim:** Effectiveness of an intervention for increasing social inclusiveness in schools
- **Subjects:** Children aged 13 to 14
- **Intervention:** Delivered in schools
- **Outcomes:** Heavy substance use
- **Randomisation clusters:** 11 intervention schools and 14 control schools
- **Observation units:** 2579 children (average of 103 per cluster)

Example: Gatehouse Study

- comparison of heavy substance use (dichotomous outcome) between trial arms
- random effects logistic regression

$$\text{logit}(\pi_{ij}) = \beta_0 + \beta_1 * G_i + u_i$$

- $y_{ij} \sim \text{Bin}(\pi_{ij}, 1)$ and coded 1 for substance use and 0 otherwise
- $u_i \sim N(0, \sigma_u^2)$
- β_1 – log odds ratio in intervention to control arm

GATEHOUSE study – Engaged in heavy substance use

Method	Odds ratio	95% CI	p-value
---------------	-------------------	---------------	----------------

Ordinary logistic regression	0.73	0.56 to 0.95	0.02
------------------------------	------	--------------	------

Random effects logistic reg. maximum likelihood (ML)	0.77	0.41 to 1.44	0.41
---	------	--------------	------

Random effects logistic reg. penalized quasi-likelihood (PQL)	0.76	0.39 to 1.48	0.43
--	------	--------------	------

CI – confidence interval

GATEHOUSE study – Engaged in heavy substance use

Method	Odds ratio	95% CI	p-value
--------	------------	--------	---------

Random effects logistic reg. - PQL (equal between-cluster variance in each arm)	0.76	0.39 to 1.48	0.43
---	------	--------------	------

$$\sigma^2_u = 0.53$$

Random effects logistic reg. – PQL (unequal between-cluster variance in control (C) and intervention (I) arms)	0.69	0.34 to 1.40	0.30
--	------	--------------	------

$$\sigma^2_{uC} = 0.32 \text{ and } \sigma^2_{uI} = 0.85$$

CI – confidence interval

Logistic regression analysis of dichotomous outcomes in CRTs

- logistic regression approaches:
 - population-averaged (PA) odds ratios
 - effect of the intervention in the population
 - e.g. marginal models using Generalised Estimating Equations
 - cluster-specific (CS) odds ratios
 - effect of the intervention conditional on cluster membership
 - e.g. random effects (“multilevel”) models
- PA odds ratio generally more appropriate in trials
 - usually interested in effects in the population
 - the CS odds ratio is not “observed” in a cluster randomised trial

Other “clustered” design types

- Multilevel models may also be applied to:
 - surveys that use cluster sampling at first stage
 - impact on confidence intervals depends on whether variables being related vary within clusters
 - multi-centre trials where individuals are randomised, stratified by centres
 - unlike CRTs, confidence intervals will generally be narrower and p-values smaller as efficiency is gained from stratifying

Population pharmacokinetic studies and nonlinear mixed effects models

*Dr Julie Simpson
Centre for MEGA Epidemiology
School of Population Health
University of Melbourne*



Outline of talk

1. Pharmacokinetics
2. Population pharmacokinetic studies
3. Nonlinear mixed effects modelling
4. Application of nonlinear mixed effects modelling to antimalarial PK studies
5. Tips for analysing data using nonlinear mixed effects modelling

Pharmacokinetics

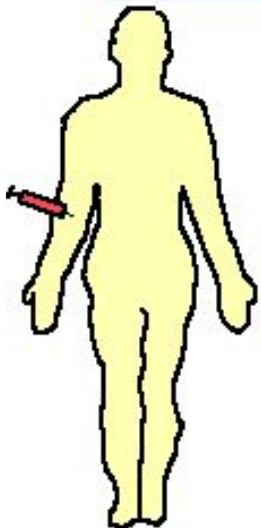
A - absorption

D - distribution

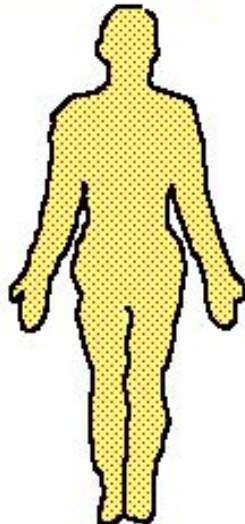
M - metabolism

E - elimination

One Compartment Model



Before
Administration

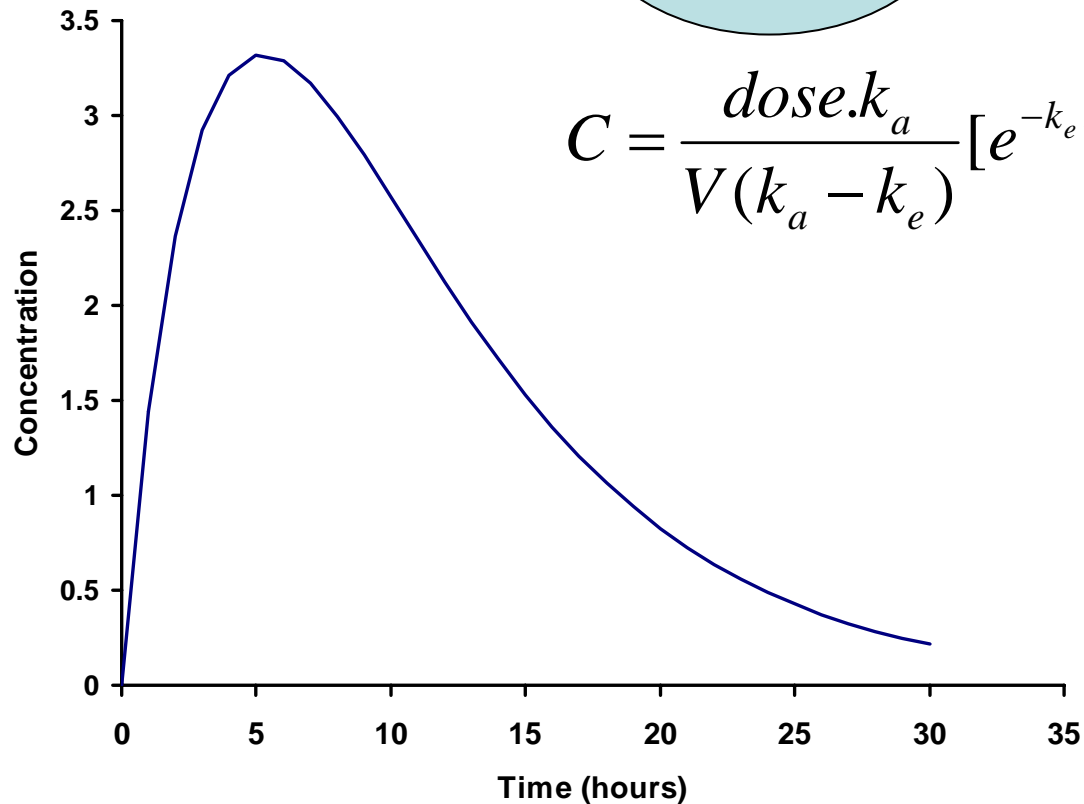
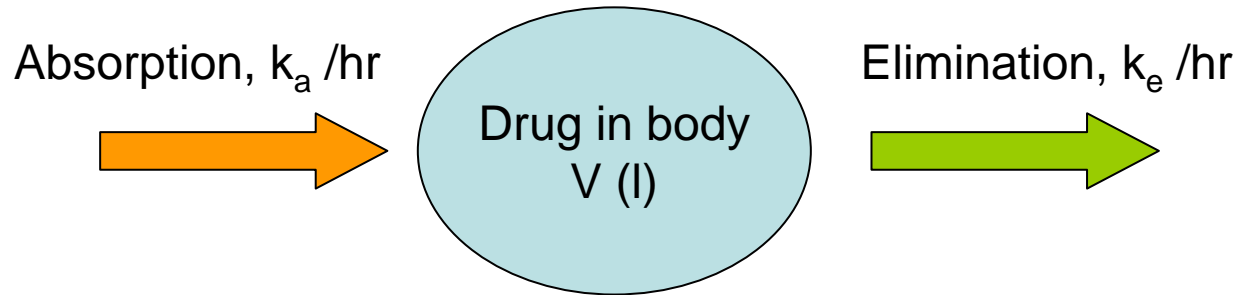


After
Administration

Why study the pharmacokinetics of a drug?

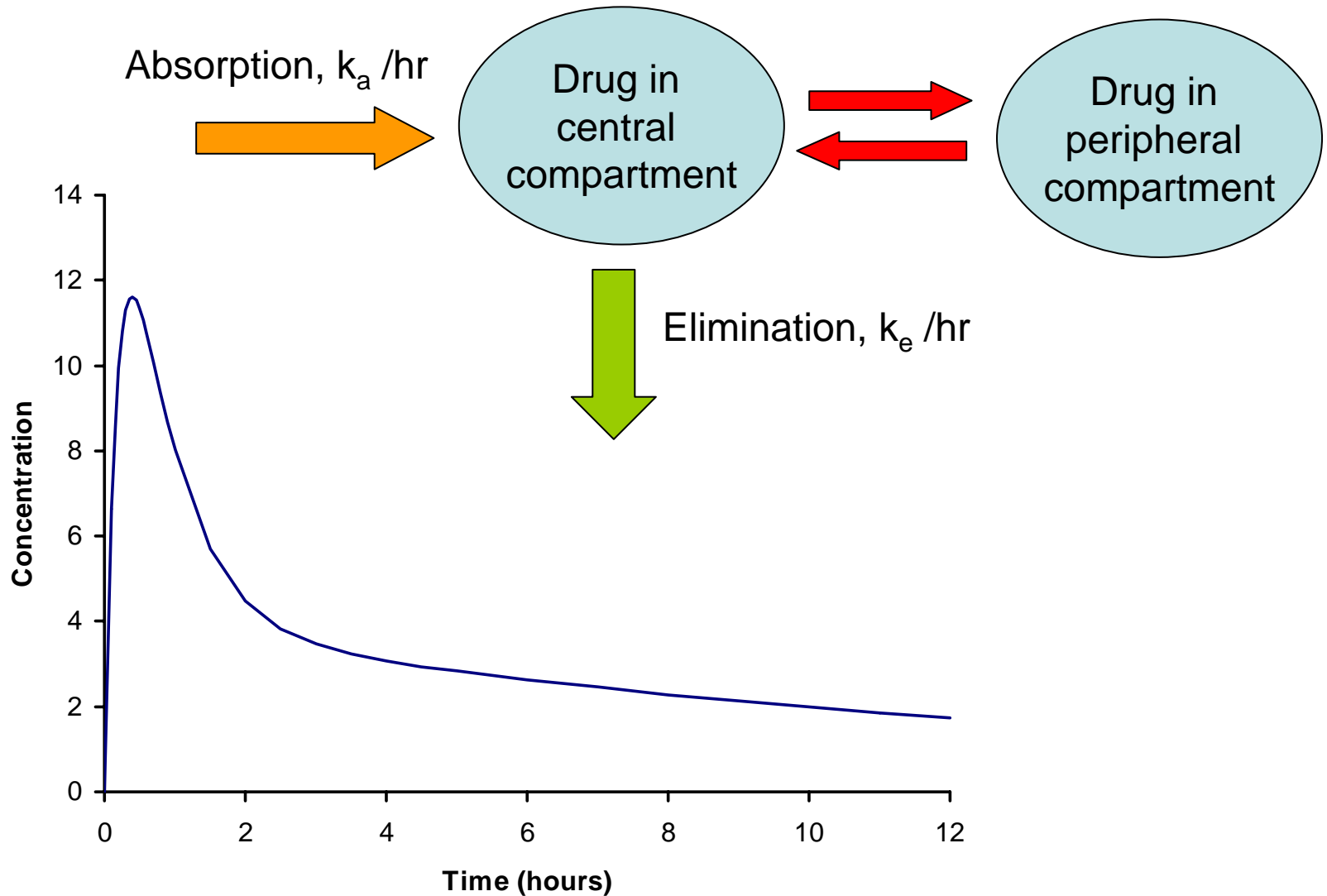
- Quantify dose-concentration relationship
- Optimise drug use
- Important part of drug development process

Pharmacokinetics



$$C = \frac{dose.k_a}{V(k_a - k_e)} [e^{-k_e.t} - e^{-k_a.t}]$$

Pharmacokinetics



Pharmacokinetic model

1. Mechanistic model
2. Model parameters have a natural physical interpretation
 k_a – absorption rate constant
3. Drug concentration depends nonlinearly on the model parameters

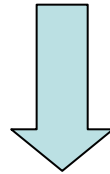
Population Pharmacokinetic Studies

1. Determine the drug concentration-time profile in the target patient population
2. Determine the patient factors that cause changes in the drug concentration-time profile

Population Pharmacokinetic Studies

Target patient population

Young children, elderly, pregnant women, very ill patients



Not ethical to perform intensive sampling



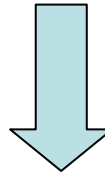
Population Pharmacokinetic Studies

Data available

Sparse data

OR

Combination of sparse and dense data

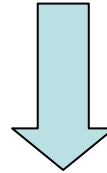


Unbalanced design

Patient ID	2 hrs post dose	4 hrs	8hrs	12 hrs	18 hrs	24 hrs	48 hrs	72 hrs
1	-	√	-	-	-	-	√	√
2	√	√	√	√	√	√	√	√
3	√	√	-	√	√	√	-	-
4	-	-	-	-	-	√	√	√

Population Pharmacokinetic Studies

How can we analyse all of the patient data?



Nonlinear mixed effects modelling

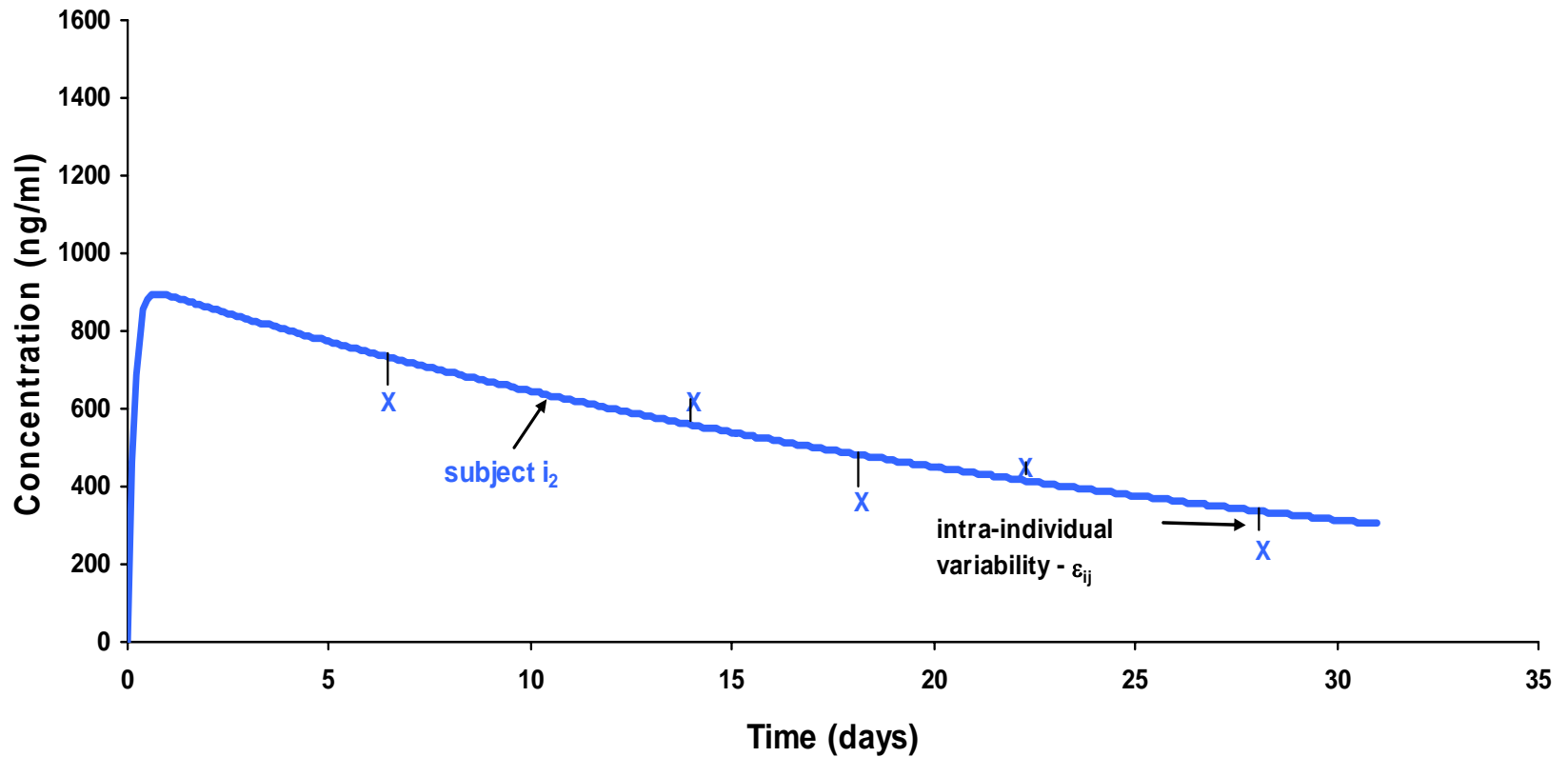


**Drug concentration depends
nonlinearly on the model parameters**

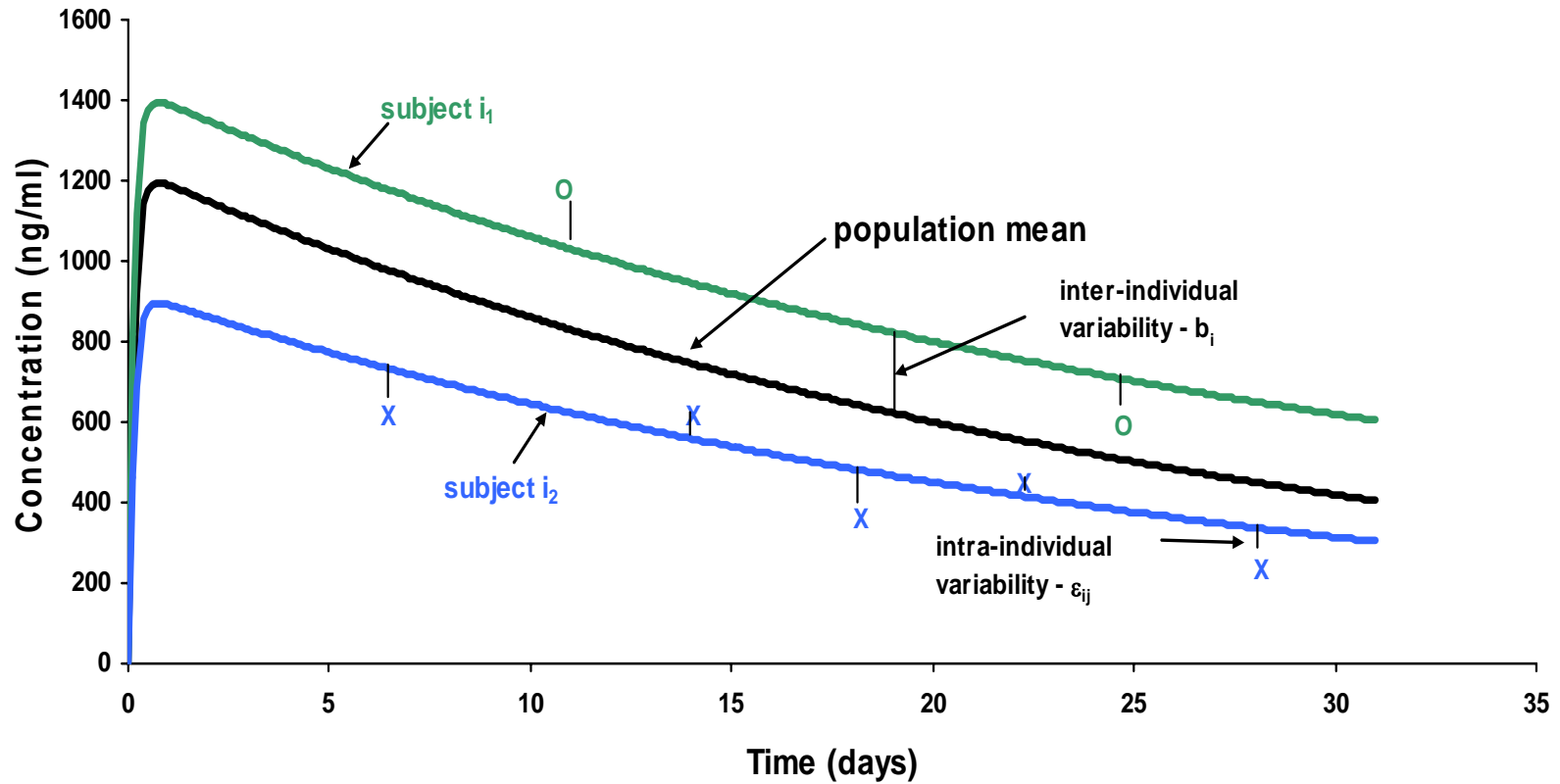


**Data for all individuals
modelled simultaneously
Fixed & random effects**

Nonlinear mixed effects modelling



Nonlinear mixed effects modelling



Nonlinear mixed effects modelling

$$C_{ij} = f(x_{ij}, \beta_i) + \varepsilon_{ij}$$

Intra-individual
error

C_{ij} – represent the j^{th} concentration of the i^{th} individual

$f(x_{ij}, \beta_i)$ - a nonlinear function of some independent variables, \mathbf{x}_{ij} (e.g. time, dose) and pharmacokinetic parameters, β_i (e.g. k_a)

$$\beta_i = \beta + b_i$$

Inter-individual
error

Population mean
estimates of k_a , etc..

Nonlinear mixed effects modelling

- Extension of linear mixed effects modelling
- Allows the regression function to depend nonlinearly on both the fixed and random effects

Practical difference between nonlinear & linear mixed effects modelling:-

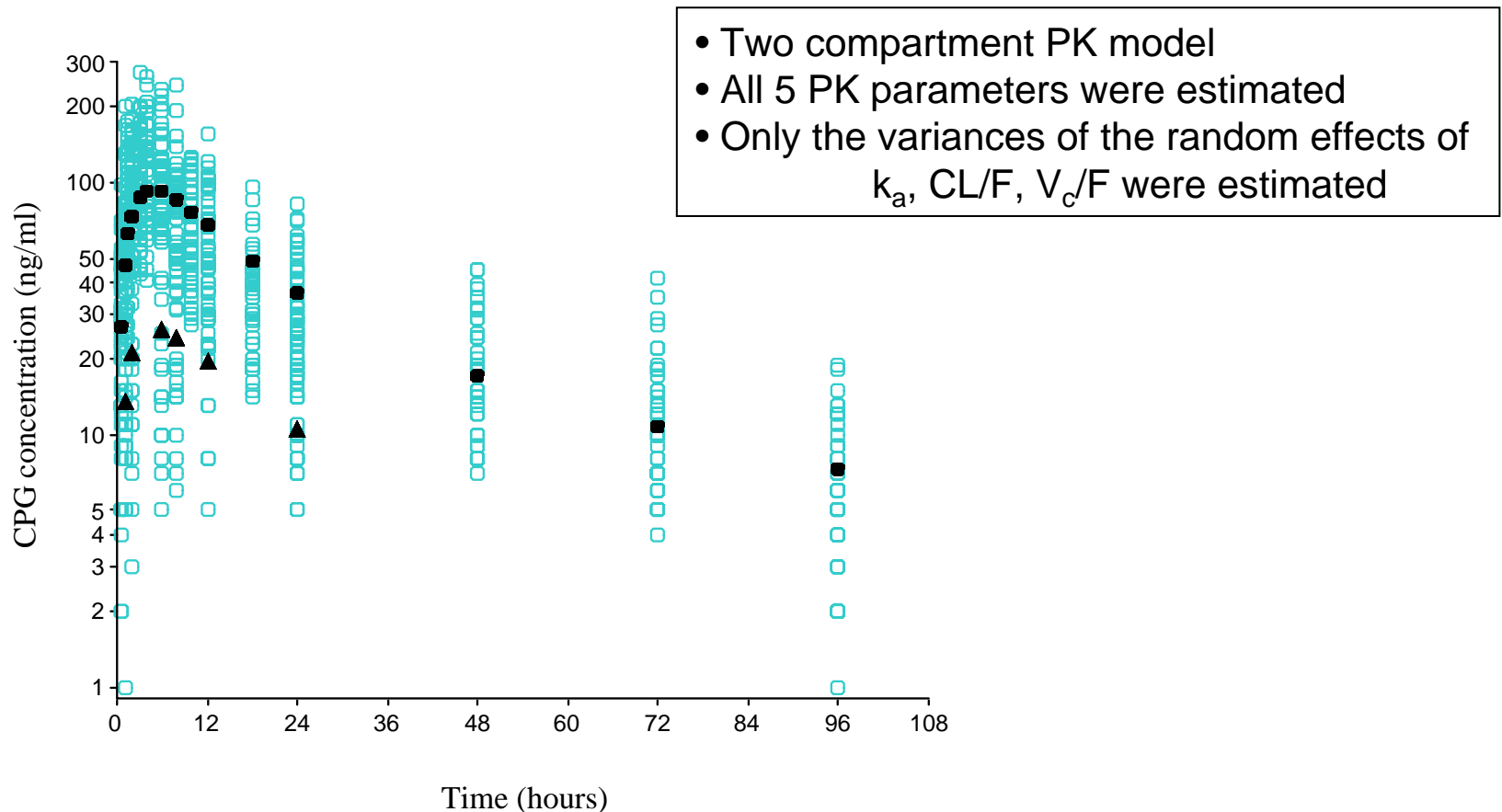
Nonlinear mixed effects modelling requires starting estimates for the fixed-effects parameters

PK models – require starting values for k_a , k_e , V , etc..

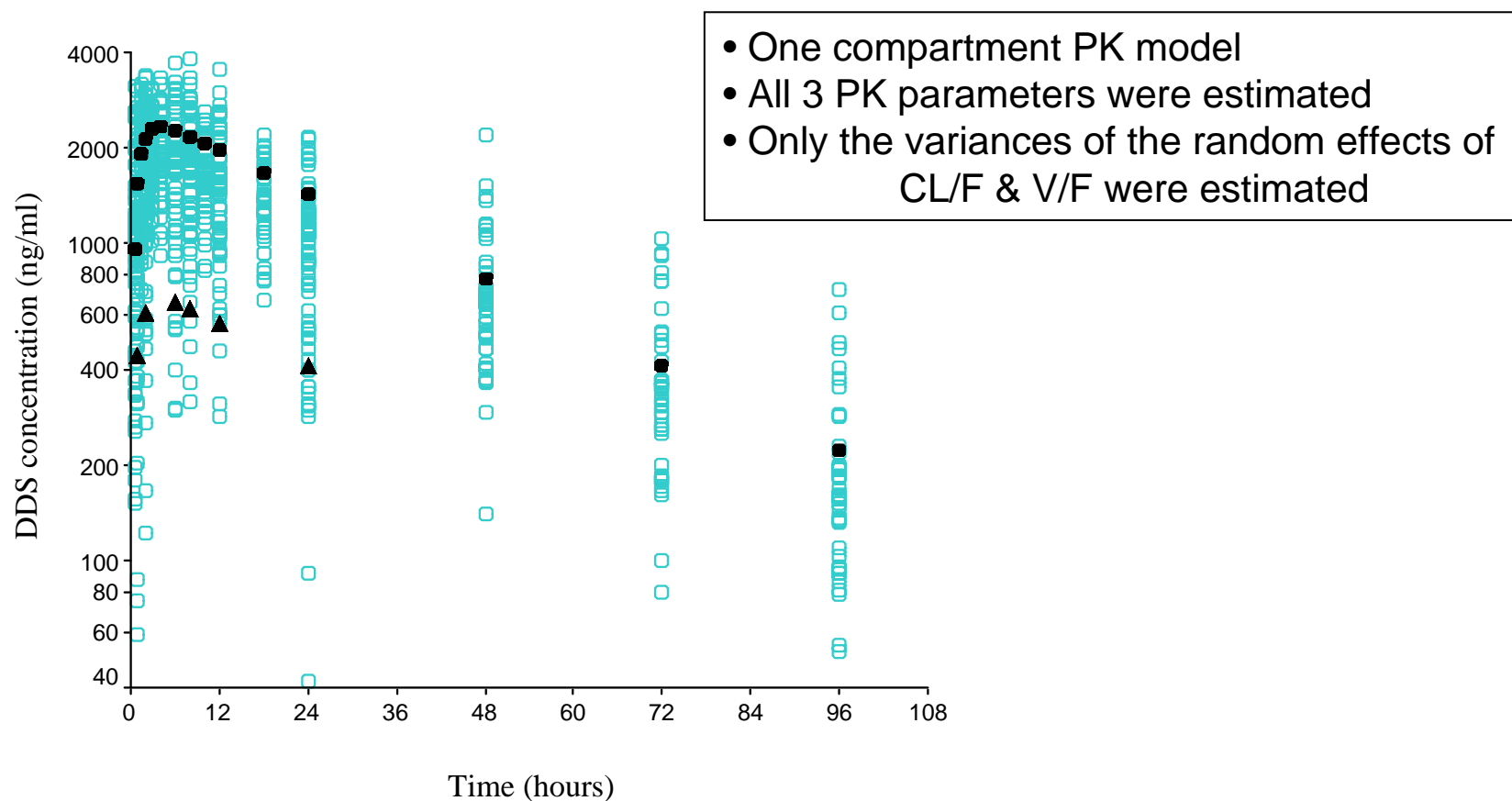
Application – Antimalarial pharmacokinetic studies

Aim:	To determine the population pharmacokinetics of chlorproguanil and dapsone
Study population:	Healthy volunteers (n=48), and adults (n=65) and children (n=68) suffering from <i>falciparum</i> malaria
Data available:	Healthy volunteers – rich sampling 0.5, 1, 1.5, 2, 3, 4, 6, 8, 10, 12, 18, 24, 48, 72, 96 hrs post dose Malaria patients – sparse sampling

Application – Population PK of chlorproguanil



Application – Population PK of dapsone



Tips for analysing data using nonlinear mixed effects modelling

1) How to determine reasonable starting estimates for the parameters in a nonlinear model?

- **Curve stripping**
- **Review of literature**
- **Simulation**

Tips for analysing data using nonlinear mixed effects modelling

2) What to do when the model won't converge?

- **Try different starting values**
- **Assume a diagonal variance-covariance matrix for the random effects**
- **Treat a parameter as fixed across all individuals**
- **Fix some of the parameters by assigning a value based on previous reports of the literature**

Tips for analysing data using nonlinear mixed effects modelling

3) How to ensure the posterior individual estimates of the biological parameters only take positive values?

- Re-parameterize the model in terms of the logarithm of the PK parameters

e.g. In the PK model replace k_a with $\exp(\beta)$

$$\beta_i = \beta + b_i, \quad \text{where } \beta = \log_e k_a$$

- Use multiplicative error models for the random effects of the PK parameters

Bivariate mixed models for assessing change: An example



Andrew Forbes

Monash University

Outline



- The practical problem: assessing simultaneous rates of change in cartilage volume and bone area in knee osteoarthritis
- A “bivariate two-level” model
- Results
- Comments

The question: Osteoarthritis of the knee

- Osteoarthritis of the knee is a disease involving the whole knee joint, especially cartilage volume and tibial bone area. Little is known about its origins.
- Current observations:
 - Cartilage volume decreases with age
 - Tibial bone area increases with age

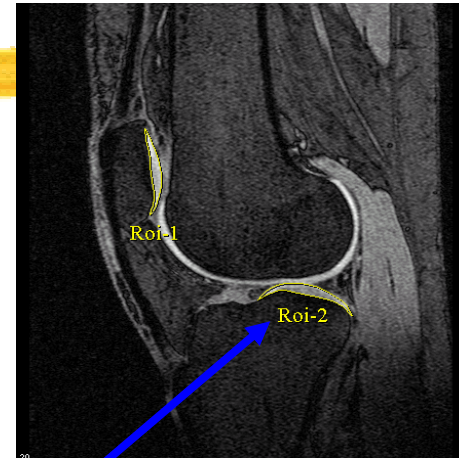
Exploratory “bivariate” questions:

Among people of same age and sex:

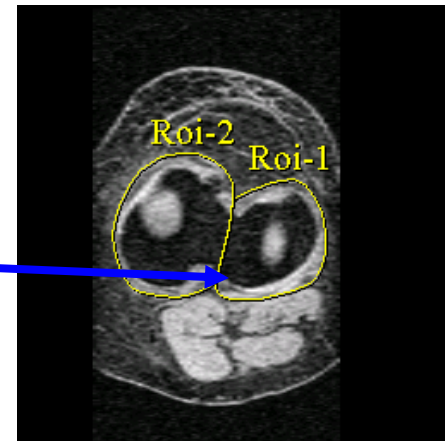
- Do people lose cartilage and get increased bone area simultaneously (ie –ve correlation of changes?)
- Is initial cartilage volume (-ve) correlated with change in bone area, or vice versa?

The data

- Community based cohort of persons with mild osteoarthritis of the knee
- MRI measurements of one affected knee taken at recruitment ($t=0$), $t\sim 2$ years, and $t\sim 4-5$ years
- Assessed knee **cartilage volume** from vertical slices
- tibial **bone area** from horizontal slices
- 73 subjects (ongoing)
- *Measurements (cart/bone) within occasions, within individuals*



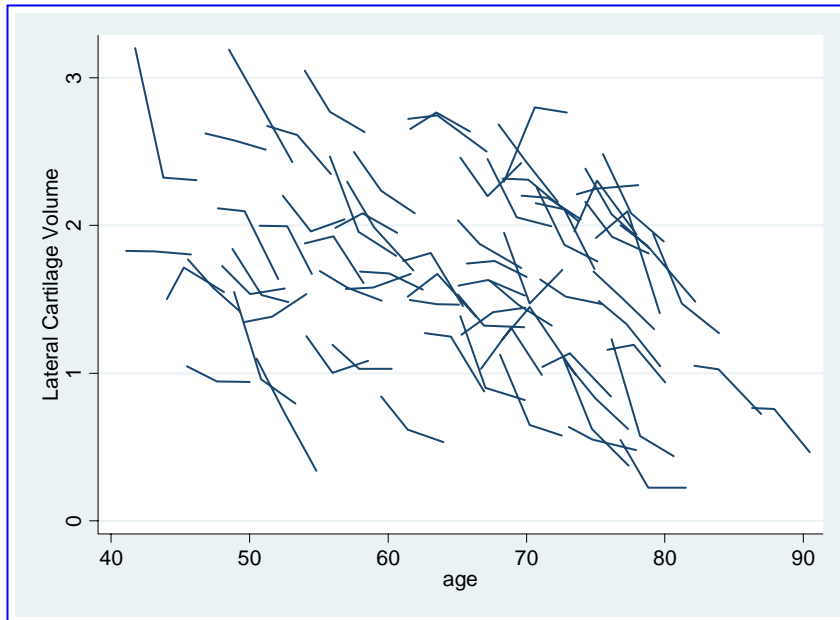
Cartilage Volume



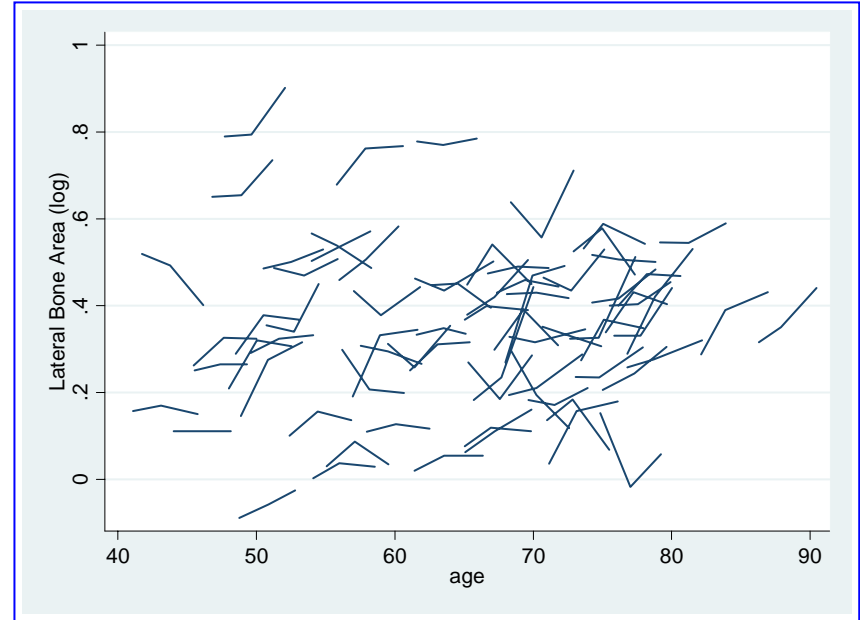
Tibial Plateau Area

The data ... (lateral only)

Cartilage volume vs age



(log) Bone area vs age



- Linear change plausible - measurement error rife!
- Longitudinal changes larger than cross sectional

Just use least squares??

- Fit straight lines using the 3 data points for each individual for cartilage and bone → intercepts and slopes (rates of change)
- Calculate average rates of change, correlate intercepts and rates of change

Is this OK?

- Reasonable approach for estimation of *fixed effects* (eg mean rates of change) when measurement times approx balanced
- However, measurement error causes:
 - *attenuation* of correlations b/w observed rates of change of bone and cartilage (since observed \sim true + error)
 - ★ Note: if cart/bone errors correlated then bias either direction
 - *Regression to the mean* for correlation of observed change with initial value

A bivariate model - concept

- Assume a *linear* change in the “true” values for each person for both cartilage volume and (log) bone area that would be observed if there was no measurement error
 - Obtain initial values and rates of change for each person
 - “Adjust” these for age and sex
 - = “age/sex adjusted random effects” for cartilage and bone for each individual (2 initial values, 2 rates of change)
- Allow these 4 random effects to be correlated + normal distn
- Independent measurement errors, different variances
 - due to independent measurement processes – different slices, observers, occasions
- *Interest is in age/sex adjusted correlations between true initial values and true rates of change of cartilage and bone*
 - Can fit this model using care with standard stats software

Notationally – 3 level MLM

i=subject, j=occasion, k=measurement (1=cart, 2=bone)

$$Y_{ijk} = \alpha_{ik} + \beta_{ik} \text{time}_{ij} + \varepsilon_{ijk}$$

Intercept \longrightarrow $\alpha_{ik} = \alpha_{0k} + \alpha_{1k} \text{sex}_i + \alpha_{2k} \text{age}_i + U_{ik}$

Slope \longrightarrow $\beta_{ik} = \beta_{0k} + \beta_{1k} \text{sex}_i + \beta_{2k} \text{age}_i + V_{ik}$

- Allow U's and V's to be correlated within and across k (measures)
 - $\text{Corr}(U_{i1}, U_{i2})$: allows initial bone and initial cartilage to be correlated
 - $\text{Corr}(V_{i1}, V_{i2})$: allows change in bone and cartilage to be correlated
 - $\text{Corr}(U_{ik}, V_{ik*})$: allows initial value and change to be correlated
- ε_{ijk} 's uncorrelated, variance σ^2_k
- Can fit using care with standard stats software

Bivariate mixed model - results

Age and sex-adjusted correlations b/w random effects (95% CI)

Cart Initial			
.01 (-.25, .25)	Bone Initial		
.11 (-.30, .48)	-.29 (-.60, .11)	Cart Slope	
-.33 (-.62, .04)	.05 (-.30, .39)	-.01 (-.49, .47)	Bone Slope

③ more initial bone → greater cartilage loss

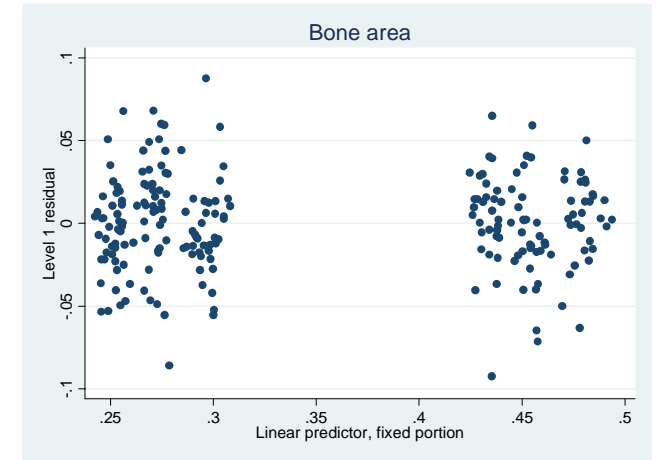
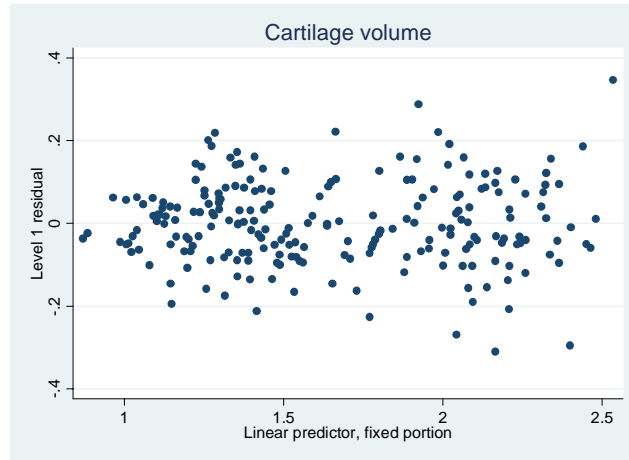
② less initial cartilage → greater increase in bone area

① No evidence of correlation between “true” rates of change of cartilage and bone!

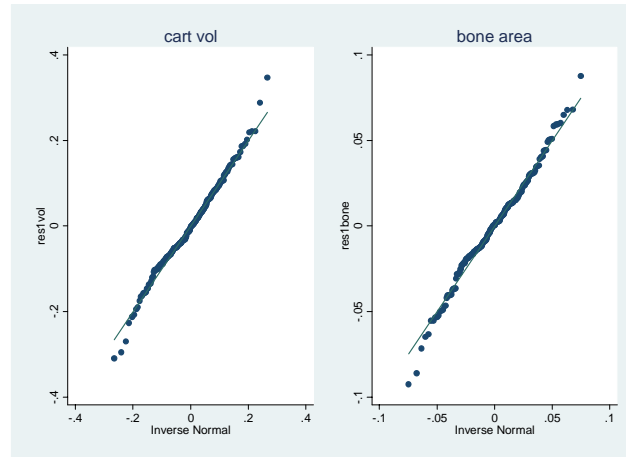
- ... clinical significance currently being pondered !
- Wide CI's due to large measurement error and only 3 times of measurement

Diagnostics

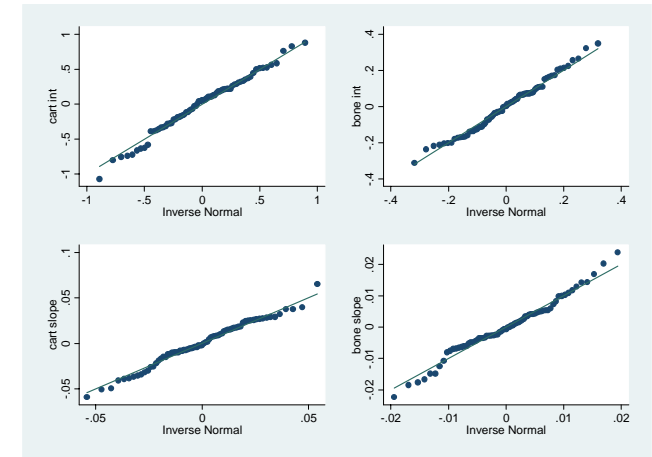
Residuals vs
predicted
means →



Normal
probability
plots →



Residuals



Predicted random effects

Extended model



- Previously assumed measurement errors (ε 's) in cartilage were uncorrelated with bone errors
 - Assess: Residuals correlated -0.10
 - Omitted time dependent covariates: Cart +, Bone - ??
- Fit correlated errors model
 - needs a lot of care in some software, easier in SAS / MLWin
 - LR test, 1df, $p=0.10$
 - Little difference in results

Conclusions



- Bivariate “growth” models can address questions concerning simultaneous change in the “true values” of two processes (ie free from the effects of measurement error)
- Model assumptions need to be checked
- Can extend to greater dimensions
 - but covariance parameters proliferate
- Standard software can be tweaked with care

Software for multilevel modelling

Lyle C Gurrin

**Centre for Molecular, Environmental, Genetic and
Analytic Epidemiology**

School of Population Health

University of Melbourne

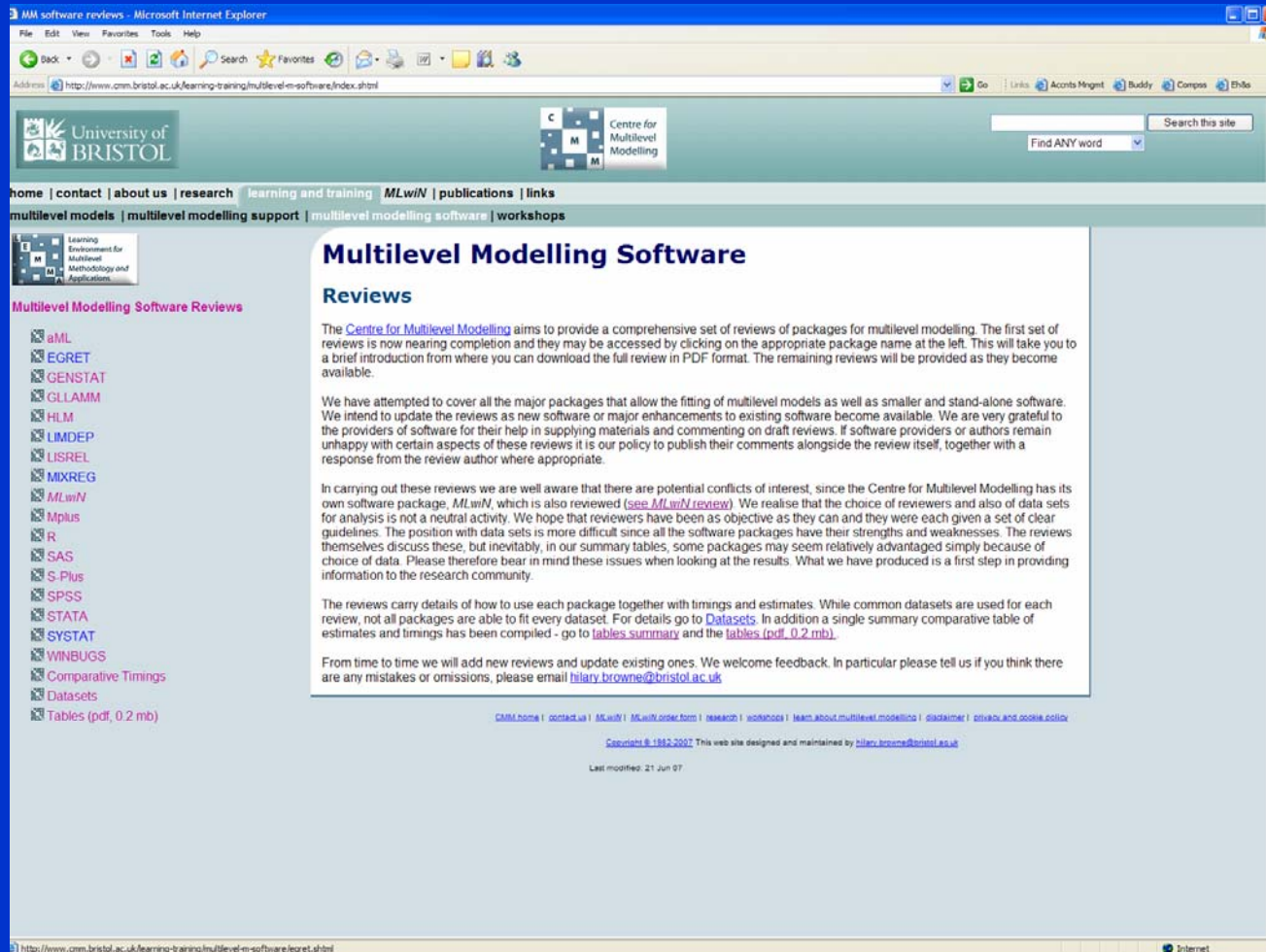


Outline

- There is a blinding array of software for multilevel modelling
 - Specialised packages like MLwiN
 - Procedures within integrated packages like SAS
- This presentation discusses the details of four programmable software packages that provide a comprehensive facility to fit and evaluate MLMs

Software reviews

The Centre for Multilevel Modelling
(www.cmm.bristol.ac.uk) has software reviews



MLM packages

aML

EGRET

GenStat

gllamm (Stata)

HLM

LIMDEP

LISREL

MIXREG

MLwiN

MPlus

R

SAS

S-Plus

SPSS

Stata

Systat

WinBUGS

MLM packages

aML

EGRET

GenStat

gllamm (Stata)

HLM

LIMDEP

LISREL

MIXREG

MLwiN

MPlus

R

SAS

S-Plus

SPSS

Stata

Systat

WinBUGS



Superior software that gives you
THE POWER TO KNOW.

- Originally offered random effects analysis of variance through **PROC VARCOMP** and **PROC GLM**
- **PROC MIXED** combines the facility of both and offers an integrated procedure:
 - Easy to handle categorical variables
 - Arbitrary number of levels
 - Complex covariance structures inc. level 1
 - Cross-classified random effects
 - Choice of estimation methods



Superior software that gives you
THE POWER TO KNOW.

- **PROC NL MIXED**

- Can fit multilevel models where the parameters appear non-linearly in the outcome-exposure relationship BUT...
- Only two levels permitted
- Categorical variables requires indicators
- Care required when choosing starting values of parameters

- **NLIN MIX** and **GLIM MIX** macros can be used to fit generalised linear mixed models



- **stata** has a suite of “**xt**” functions
 - Analyse panel/longitudinal data using random effects
 - Two levels with random intercepts only
- Version 9 introduced **xtmixed** for linear mixed models of continuous outcomes
 - Multiple levels of nesting
 - Random coefficients
 - Cross classified random effects
 - Complex covariance structures at each level



- **gllamm** (generalised linear latent and mixed models) fits multilevel models for different types of responses
 - up to three levels models
 - random coefficients for normal and discrete data using Gaussian quadrature (www.gllamm.org)
- Version 10 introduced **xtmelogit** and **xtmepoisson** which fit MLMs for binary and count data
- User developed commands
 - **metareg** (meta-analysis regression)
 - **rpoisson** (Poisson regression with a random effect)
 - **reoprobit** (random effects ordered probit)



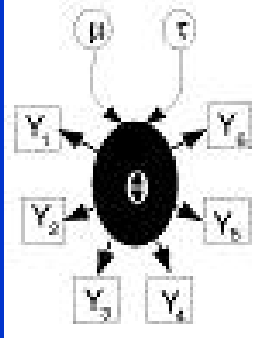
The R Project for Statistical Computing

- The premier MLM package in R is `nlme` by Jose Pinheiro and Douglas Bates
 - `lme` for linear mixed models
 - `nlme` for non-linear mixed effects models (assumes random effects normal)
- The `lme4` package (with flagship `lmer`) simplifies the syntax for specifying random effects.
- All offer multiple levels, cross-classified random effects and complex covariance structures

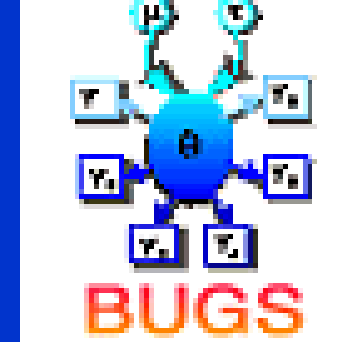


The R Project for Statistical Computing

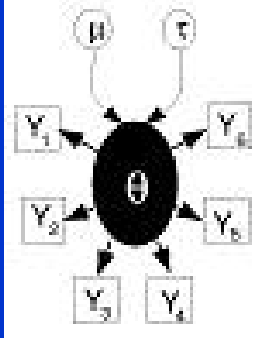
- There are several user contributed packages in `R` for generalised linear mixed models:
 - `glmmML` (maximum likelihood)
 - `glmmPQL` (penalised quasi-likelihood, from a package by Bill Venables and Brian Ripley)
 - `glmmGibbs` (Gibbs sampler MCMC approach)



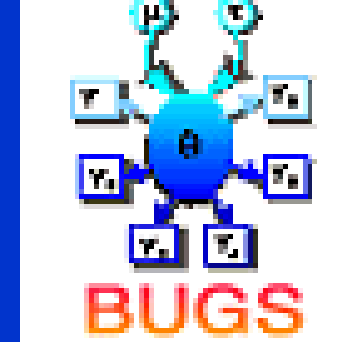
The BUGS project



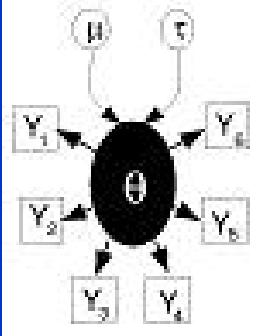
- WinBUGS was developed by British MRC BSU in Cambridge, joint now with Imperial College London.
- Version for Linux and MAC (OpenBUGS) and \mathbf{R} (BRugs) are also available.
- BUGS uses Markov chain Monte Carlo (MCMC) methods, so it fits models by iterative simulation.



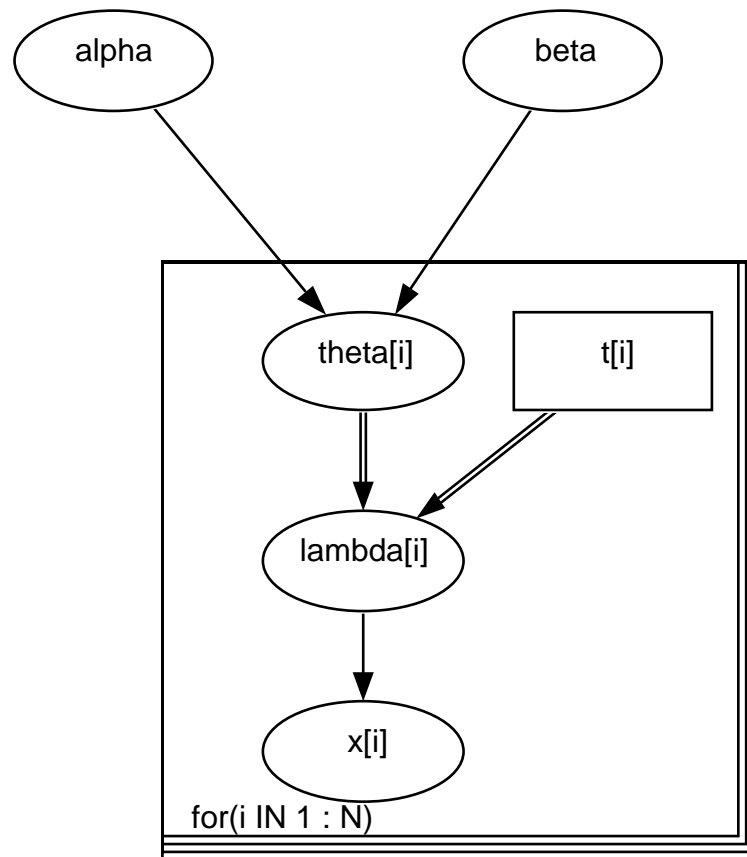
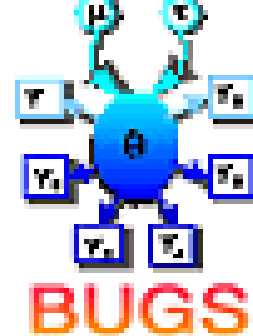
The BUGS project

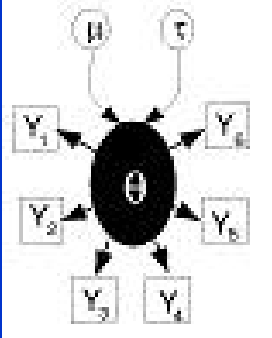


- BUGS represents statistical models as graphs
 - parameters and data are *nodes*
 - missing *edges* represent conditional independence.
- Multilevel models are easily represented using a graph (and hence in BUGS) due to hierarchical structure and conditional independence assumptions.

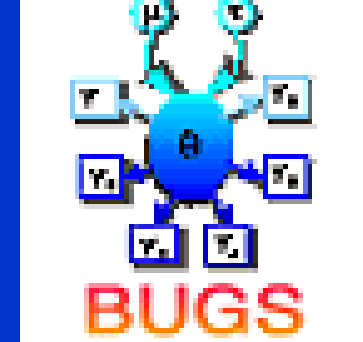


The BUGS project





The BUGS project



Advantages

- Can fit models that appear simple but for which there is no analytic solutions and standard approximations fail.
- Inference can be performed for any function of the parameters.
- MLwiN can generate some WinBUGS model code.

Disadvantages

- Time consuming
- Full probability specification requires prior probability distributions for the model parameters
- The embedded statistical theory is challenging

Final Comments

- A formal comparison of MLM software is available on the CMM website
- Mainstream statistical packages now have flexible MLM routines, which presents a challenge for specialised software
- Many of the authors of MLM software have written books on the topic – too numerous to list here!

Advanced applications in multilevel models

Alastair H Leyland

MRC Social and Public Health Sciences Unit
Glasgow, Scotland

Øyvind Næss

National Institute of Public Health
Oslo, Norway

Air pollution, social deprivation and mortality

- Exposure to air pollution has been shown to be associated with increased mortality
- Distribution of air pollution is not equitable
- To what extent does social deprivation explain the effect of neighbourhood-level air pollution on mortality?
- Does this depend on the deprivation measure used, and whether it is an individual or contextual measure?

Data and methods

- Cohort of all inhabitants of Oslo aged 50-74 in 1992
- Restricted to those who had lived at same residence in 1980 and 1992: 27,943M & 38,832F
- Deaths recorded 1992-98
- Concentration of pollutants ($PM_{2.5}$) estimated in 468 neighbourhoods based on hourly data 1992-95
- Person weighted median concentrations were used to calculate average exposure quartiles for each neighbourhood

Indicators of deprivation

- 1990
 - ▲ Primary education only
 - ▲ Equivalised household income below median
- 1980 Census
 - ▲ Manual occupational class
 - ▲ Housing tenure (does not own own dwelling)
 - ▲ Type of dwelling (flat)
 - ▲ Overcrowding (more than one person per room)
- Individual measures also aggregated to neighbourhood level and standardised

Multilevel spatial model for mortality

$$y_{ij} \sim \text{Bin}(1, \pi_{ij})$$

$$\text{logit}(\pi_{ij}) = \mathbf{X}\boldsymbol{\beta} + v_j + u_j$$

$$v_j \mid v_{-j} \sim N(\bar{v}_j, \sigma_v^2 / n_j)$$

$$u_j \sim N(0, \sigma_u^2)$$

Age +

- 1) Individual deprivation
- 2) Area deprivation
- 3) Individual deprivation and air pollution
- 4) Area deprivation and air pollution
- 5) Individual and area deprivation and air pollution

Results: ORs associated with education, men 50-74

	Ind	Area	PM _{2.5}
Ind	1.41 (1.31-1.52)		
Area		1.30 (1.24-1.36)	
Ind+PM	1.41 (1.31-1.52)		1.10 (1.03-1.17)
Area+PM		1.28 (1.22-1.35)	1.05 (1.00-1.11)
Ind+Area+PM	1.34 (1.24-1.43)	1.22 (1.16-1.28)	1.06 (1.00-1.11)

Results: variances associated with education, men 50-74

	Spatial	Heterogeneity
Ind	0.158 (0.093-0.237)	0.002 (0.000-0.012)
Area	0.077 (0.038-0.134)	0.003 (0.000-0.012)
Ind+PM	0.133 (0.069-0.210)	0.003 (0.000-0.014)
Area+PM	0.062 (0.022-0.116)	0.003 (0.000-0.015)
Ind+Area+PM	0.061 (0.022-0.116)	0.003 (0.000-0.014)

Results: ORs associated with income, men 50-74

	Ind	Area	PM _{2.5}
Ind	1.49 (1.39-1.58)		
Area		1.22 (1.17-1.28)	
Ind+PM	1.48 (1.39-1.58)		1.09 (1.02-1.17)
Area+PM		1.22 (1.16-1.27)	1.05 (1.00-1.11)
Ind+Area+PM	1.44 (1.35-1.53)	1.16 (1.11-1.21)	1.05 (1.00-1.12)

Results: ORs associated with education, women 50-74

	Ind	Area	PM _{2.5}
Ind	1.40 (1.30-1.50)		
Area		1.26 (1.20-1.32)	
Ind+PM	1.40 (1.30-1.50)		1.10 (1.04-1.17)
Area+PM		1.24 (1.19-1.30)	1.05 (1.00-1.10)
Ind+Area+PM	1.32 (1.23-1.42)	1.18 (1.12-1.24)	1.05 (1.00-1.11)

Conclusions

- Effect of air pollution largely uninfluenced by individual deprivation measures
- Air pollution correlated with, and partly explained by, area deprivation measures
- Area deprivation effect unchanged when adjusted for air pollution
- Næss Ø, Piro FN, Nafstad P, Davey Smith G, Leyland AH. Air pollution, social deprivation and mortality. A multilevel cohort study of 468 small neighborhoods in Oslo, Norway. *Epidemiology*, in press.

Longitudinal effect of context over the life course on health

- Effect of early life (and accumulated) influences on adult health recognised
- Health is patterned by current context (neighbourhood, workplace, school etc)
- To what extent do contexts from early life influence subsequent health?
- What influence does area of residence at 4 different time points – covering a period of 30 years – have on mortality?

Data and methods

- Cohort of all inhabitants of Oslo aged 30-69 in 1990
- Restricted to those men who had lived in Oslo in 1960, 1970, 1980 and 1990 (n=49,736)
- Deaths recorded 1990-98
- Area of residence recorded at each Census
- 69 neighbourhoods were defined based on area codes used in the 1960 Census
- Analyses stratified into 10-year age cohorts

Competing models for mortality:

1 Influence of one time only

$$y_{i\{j\}} \sim \text{Bin}\left(1, \pi_{i\{j\}}\right)$$

e.g. 1960

$$\text{logit}\left(\pi_{i\{j\}}\right) = \mathbf{X}\boldsymbol{\beta} + u_{j_{60}}^{(60)}$$

$$u_j^{(60)} \sim N\left(0, \sigma_u^{(60)2}\right)$$

Fails to take into account area of residence in 1970, 1980 and 1990

Competing models for mortality:

2 Multiple membership model

$$y_{i\{j\}} \sim \text{Bin}\left(1, \pi_{i\{j\}}\right)$$

$$\text{logit}\left(\pi_{i\{j\}}\right) = \mathbf{X}\boldsymbol{\beta} + u_{j_{60}} + u_{j_{70}} + u_{j_{80}} + u_{j_{90}}$$

$$u_j \sim N\left(0, \sigma_u^2\right)$$

Assumes area effects constant over time

Assumes equal importance of area at all stages of the life course

Competing models for mortality:

3 Cross-classified model

$$y_{i\{j\}} \sim \text{Bin}\left(1, \pi_{i\{j\}}\right)$$

$$\text{logit}\left(\pi_{i\{j\}}\right) = \mathbf{X}\boldsymbol{\beta} + u_{j_{60}}^{(60)} + u_{j_{70}}^{(70)} + u_{j_{80}}^{(80)} + u_{j_{90}}^{(90)}$$

$$u_j^{(Y)} \sim N\left(0, \sigma_u^{(Y)2}\right) \quad \text{Cov}\left(u_j^{(Y_1)}, u_j^{(Y_2)}\right) = 0$$

Assumes no correlation between area effects over time

Form of covariance matrices

$$\mathbf{u} \sim N(\mathbf{0}, \Sigma)$$

Multiple membership

$$\begin{bmatrix} \sigma_u^2 & \sigma_u^2 & \sigma_u^2 & \sigma_u^2 \\ \sigma_u^2 & \sigma_u^2 & \sigma_u^2 & \sigma_u^2 \\ \sigma_u^2 & \sigma_u^2 & \sigma_u^2 & \sigma_u^2 \\ \sigma_u^2 & \sigma_u^2 & \sigma_u^2 & \sigma_u^2 \end{bmatrix}$$

Cross-classified

$$\begin{bmatrix} \sigma_u^{(60)^2} & 0 & 0 & 0 \\ 0 & \sigma_u^{(70)^2} & 0 & 0 \\ 0 & 0 & \sigma_u^{(80)^2} & 0 \\ 0 & 0 & 0 & \sigma_u^{(90)^2} \end{bmatrix}$$

Competing models for mortality:

4 Correlated cross-classified model

$$y_{i\{j\}} \sim \text{Bin}\left(1, \pi_{i\{j\}}\right)$$

$$\begin{aligned} \text{logit}\left(\pi_{i\{j\}}\right) = & \mathbf{X}\boldsymbol{\beta} + u_{j_{60}}^{(60)} + u_{j_{70}}^{(70)} + u_{j_{80}}^{(80)} + u_{j_{90}}^{(90)} \\ & + v_{j_{60}}^{(60)} + v_{j_{70}}^{(70)} + v_{j_{80}}^{(80)} + v_{j_{90}}^{(90)} \end{aligned}$$

$$u_j^{(Y)} \sim N\left(0, \sigma_u^{(Y)2}\right) \quad \text{Cov}\left(u_j^{(Y_1)}, u_j^{(Y_2)}\right) = 0$$

$$v_j^{(Y)} \sim N\left(0, \sigma_v^{(Y)2}\right) \quad \text{Cov}\left(v_j^{(Y_1)}, v_j^{(Y_2)}\right) = \sqrt{\sigma_v^{(Y_1)2} \sigma_v^{(Y_2)2}}$$

Form of covariance matrix

$$\mathbf{u} \sim N(\mathbf{0}, \Sigma)$$

Correlated cross-classified

$$\begin{bmatrix} \sigma_u^{(60)2} + \sigma_v^{(60)2} & \sqrt{\sigma_v^{(60)2} \sigma_v^{(70)2}} & \sqrt{\sigma_v^{(60)2} \sigma_v^{(80)2}} & \sqrt{\sigma_v^{(60)2} \sigma_v^{(90)2}} \\ \sqrt{\sigma_v^{(60)2} \sigma_v^{(70)2}} & \sigma_u^{(70)2} + \sigma_v^{(70)2} & \sqrt{\sigma_v^{(70)2} \sigma_v^{(80)2}} & \sqrt{\sigma_v^{(70)2} \sigma_v^{(90)2}} \\ \sqrt{\sigma_v^{(60)2} \sigma_v^{(80)2}} & \sqrt{\sigma_v^{(70)2} \sigma_v^{(80)2}} & \sigma_u^{(80)2} + \sigma_v^{(80)2} & \sqrt{\sigma_v^{(80)2} \sigma_v^{(90)2}} \\ \sqrt{\sigma_v^{(60)2} \sigma_v^{(90)2}} & \sqrt{\sigma_v^{(70)2} \sigma_v^{(90)2}} & \sqrt{\sigma_v^{(80)2} \sigma_v^{(90)2}} & \sigma_u^{(90)2} + \sigma_v^{(90)2} \end{bmatrix}$$

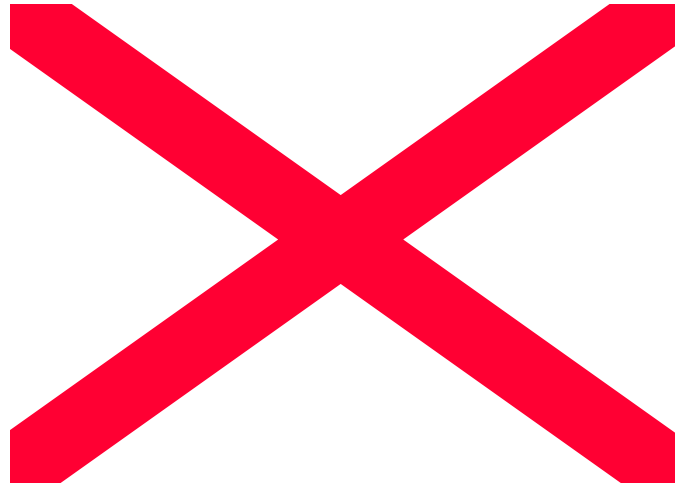
Model comparison by cohort

Cohort	<i>One year</i>	<i>MM</i>	<i>XC</i>	<i>CXC</i>
30-39	2838 ^a	2862	2840	2835
40-49	4187 ^a	4204	4184	4178
50-59	6952 ^a	6961	6954	6949
60-69	18249 ^b	18236	18246	18233

Values shown are DIC ^a1990 ^b1980

Life course models for area contributions

Contribution of residence at each decade



Conclusions

- In the youngest age groups, most recent area of residence has greatest contribution to mortality
- In the oldest age group the contribution from ages 30-9 to 60-9 is more evenly distributed
- Næss Ø, Davey Smith G, Claussen B, Leyland AH. Lifecourse influence of residential area on cause specific mortality. *J Epidemiol Community Health*, in press.

Hierarchical model of seasonality in cardiovascular disease

Multilevel Modelling Workshop

The 16th Annual Scientific Meeting of the AEA

Adrian Barnett
a.barnett@uq.edu.au

School of Population Health
The University of Queensland

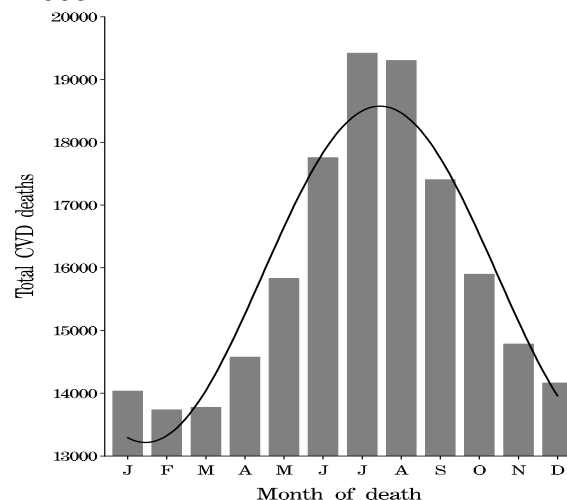
26 August 2007

Seasonality in cardiovascular disease

- Many studies around the world have shown a seasonality in cardiovascular disease (CVD)
- Increase in events and deaths in cold weather; sharp peaks in heat waves
- Multiple possible pathways for cold effect:
 - \downarrow Temperature $\Rightarrow \uparrow$ Blood pressure
 - \downarrow Sunlight $\Rightarrow \uparrow$ Blood pressure
 - Winter $\Rightarrow \uparrow$ Flu $\Rightarrow \uparrow$ Inflammation
 - Winter $\Rightarrow \downarrow$ Decreased activity $\Rightarrow \uparrow$ BMI & Cholesterol
 - \downarrow Temperature $\Rightarrow \uparrow$ Heating $\Rightarrow \uparrow$ Air pollution $\Rightarrow \uparrow$ Heart rate, inflammation & BP
- Do hierarchical models offer the chance to investigate these complex relationships?

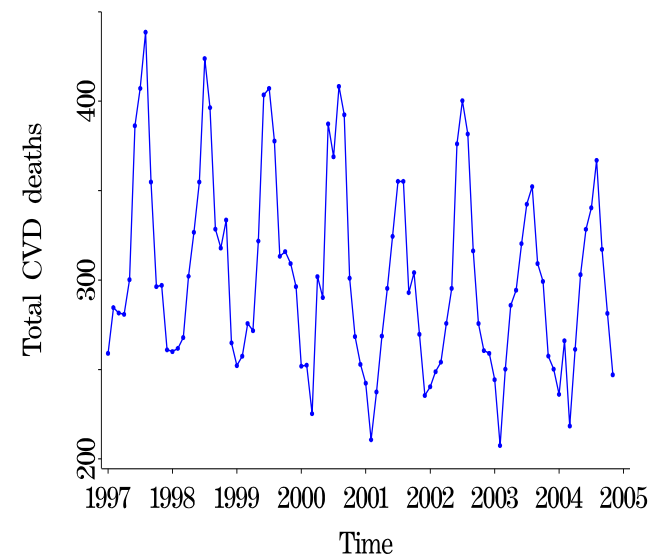
Seasonality in CVD in Australia

Total number of CVD deaths by month, 8 Australian state and territory capitals, 1997–2003.



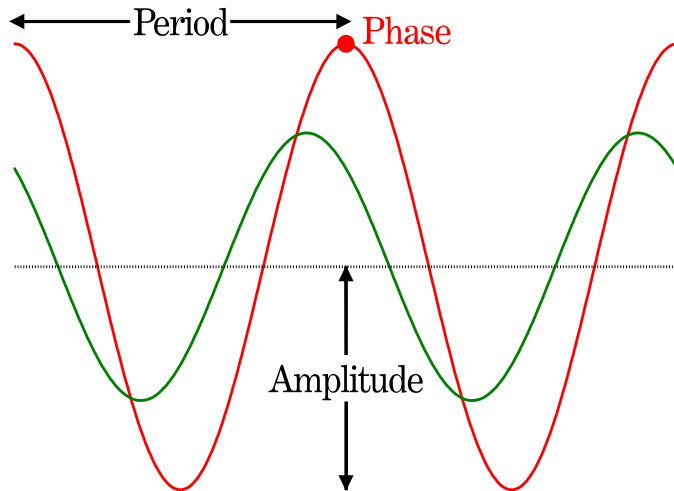
The y-axis starts at 13,000 deaths

Example time series: Sydney men, aged 65+



A sinusoidal model

$y = \text{Amplitude} \times \cos(\text{Time} \times \omega - \text{Phase})$, where $\omega = 2\pi/\text{Period}$



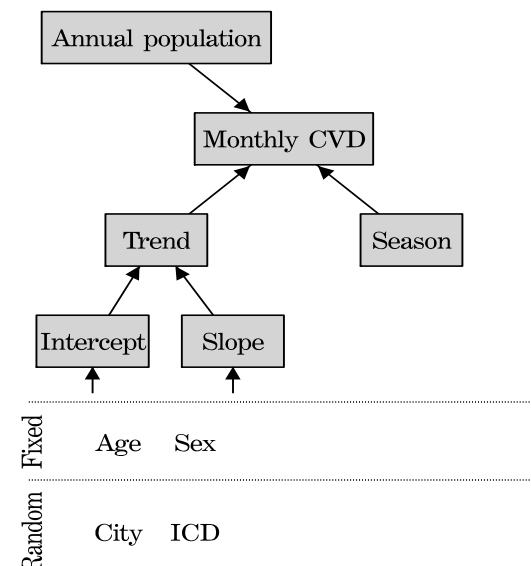
Data description

- Monthly counts of cardiovascular deaths for:
 - Year 1997 to 2004 (8 years)
 - The eight Australian states and territory capital cities
 - Age (18–64 years, 65+ years)
 - Sex
 - ICD-10 code (7 groups using the subsections of the ICD classification)
- Data from the Australian Institute of Health and Welfare.
Co-investigators: Michael de Looper (AIHW), John Fraser (Prince Charles Hospital)

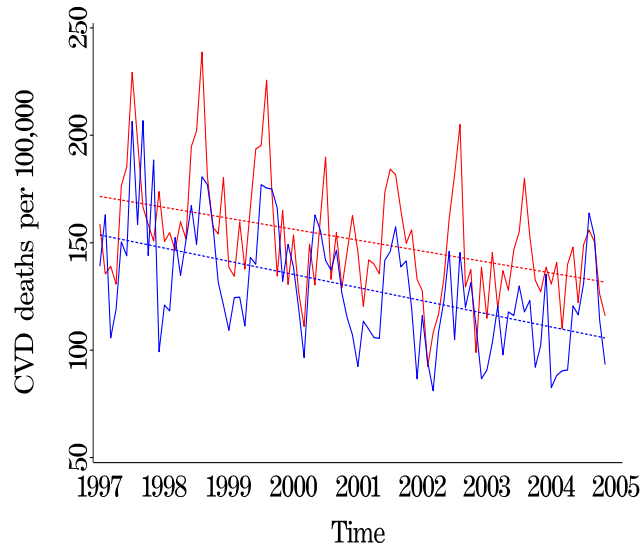
Research questions

- Is the seasonal amplitude in CVD dependent on:
 - Location *How does climate effect seasonality?*
 - Time *Are things getting worse/better?*
 - Age group *Is their greater risk with greater frailty?*
 - Sex *Is there a biological or behavioural pathway?*
 - ICD class *Clues to the pathway?*

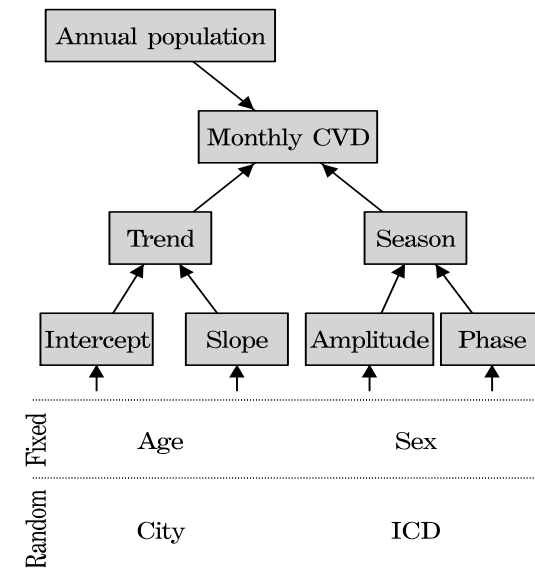
Building the hierarchical model



Intercept & slope example (Brisbane & Perth)



Building the hierarchical model



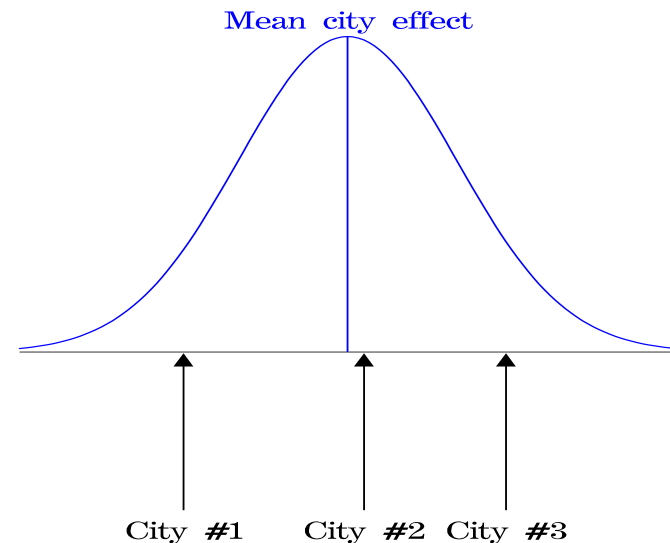
Random vs fixed effects

- Random effects are allowed to vary around a mean effect
- This variance is determined by Normal distribution

$$\alpha_j \sim N(\bar{\alpha}, \sigma_{\alpha}^2), \quad j = 1, \dots, m$$

- For example, the intercept of CVD deaths (per 100,000) is allowed to vary by city
- A fixed effect would be the same for every city (i.e., $\bar{\alpha}$)

Random vs fixed effects



Random vs fixed effects

- Random effects occur because of heterogeneity in an explanatory variable
 - e.g., the mean rate of CVD is high in one city because of a complex mix of diet, smoking, and other lifestyle factors
- Random effect might occur because the same underlying effect is altered by local conditions
 - e.g., the biological effect of cold temperatures is the same but it is modified by local housing types
- Random effects model unmeasured effects. If we knew the drivers behind the random effect these could be added to the regression model and the random difference would be explained
- Beware: changing from a fixed to a random effect can lead to a big increase in the number of parameters

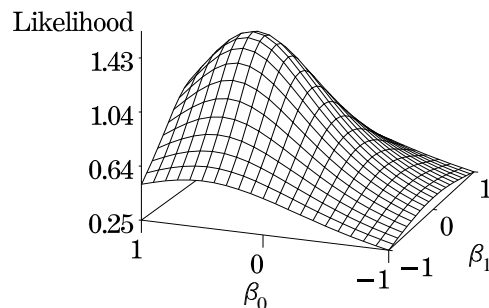
The WinBUGS software package

- WinBUGS is a popular Bayesian analysis package
- Win = Windows, BUGS = Bayesian analysis Under Gibbs Sampling
- Gibbs Sampling is a particular type of Markov chain Monte Carlo (MCMC)
- MCMC is a technique for finding solutions

Maximum likelihood

Finding solutions using classical statistics

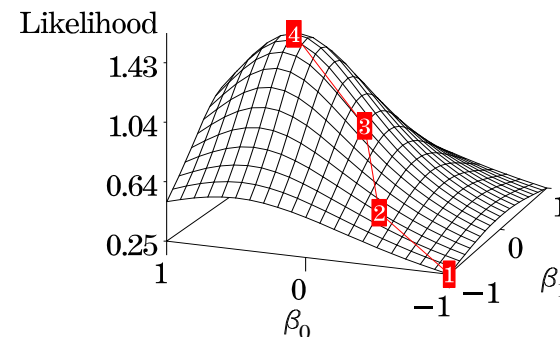
- **Likelihood** means how likely a model is given the data. Indication of model fit.
- Imagine we are fitting a linear regression with an intercept β_0 and slope β_1 . If we could picture the likelihood it might look like a hill:



Maximum likelihood

Finding solutions using classical statistics

- An iterative maximum likelihood finds the quickest way to the top given a starting point

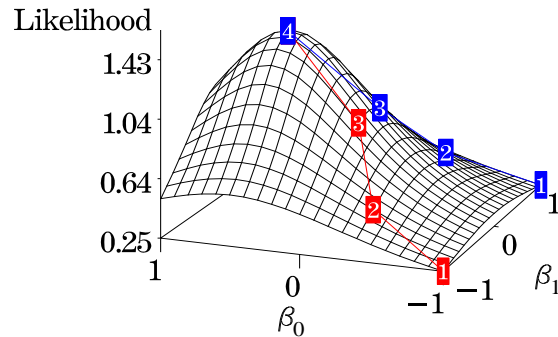


- 4 iterations to give $\hat{\beta}_0 = 0.5$, $\hat{\beta}_1 = 0$.

Maximum likelihood

Finding solutions using classical statistics

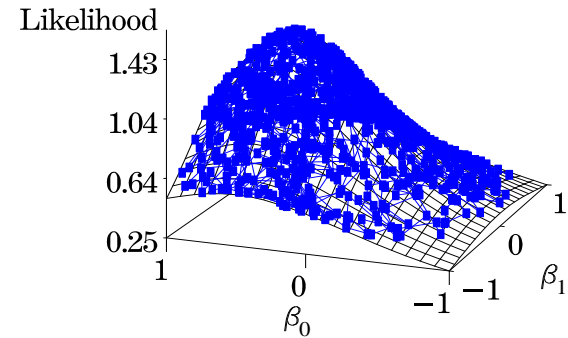
- An iterative maximum likelihood finds the quickest way to the top given a starting point



- 4 iterations to give $\hat{\beta}_0 = 0.5$, $\hat{\beta}_1 = 0$.

Markov chain Monte Carlo

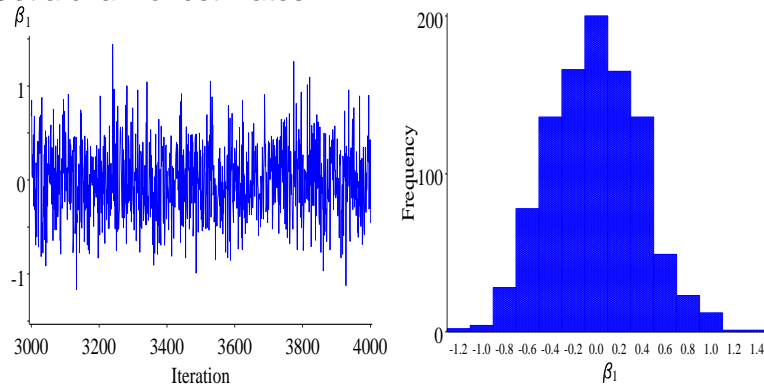
- MCMC attempts to cover the whole likelihood (not just the peak)



- 100 iterations 1000 iterations

Markov chain Monte Carlo

- Get a chain of estimates



- Make a histogram of the chain

Results

CVD deaths in Australian capital cities, 1997–2003

- Mean phase (peak time) 7.7 months, 95% posterior interval 7.4, 8.0 months
- Mean seasonal amplitudes (percent increase in late July)

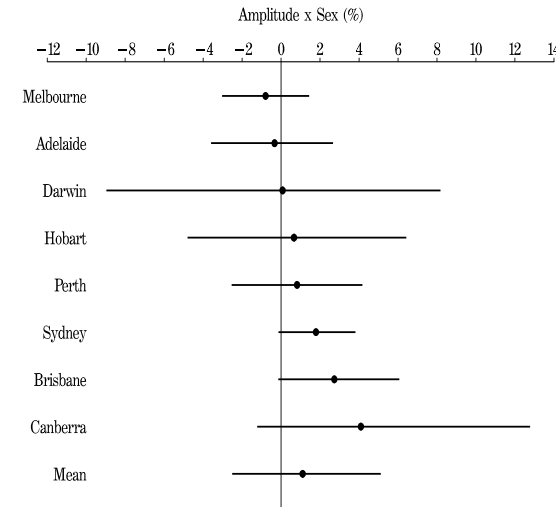
	Mean	95% posterior interval
Mean	12.9%	6.4%, 18.1%
<i>Changes due to:</i>		
Age (65+)	8.3%	0.8%, 14.8%
Time (year)	-0.5%	-0.7%, -0.2%
Sex (women)	1.2%	-2.2%, 5.2%

Results: fixed vs random sex

- A change in the effect of season by sex might indicate either:
 - Behavioural difference (e.g., clothes worn are less protective against the cold)
 - Biological difference (e.g., different temperature control: hypothalamus neurons & estrogen receptors)
- If the effect is behavioural we might expect a difference by location. A biological effect should be more consistent
- Can test this effect by comparing a fixed effect for sex to a random effect

Results: random sex effect by city

- Change in amplitude for women, mean and 95% posterior interval. Results ordered by the mean



Results: fixed vs random sex

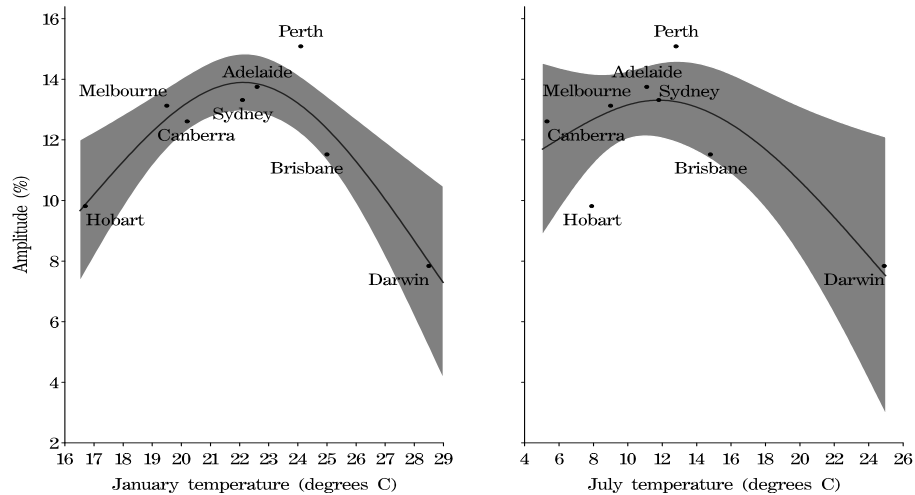
- Increase in amplitude for women:
 - Fixed effect: 0.9% (-0.4%, 2.1%)
 - Mean random effect: 1.2% (-2.2%, 5.2%)
- Deviance information criterion (DIC), Bayesian version of the Akaike Information Criterion (AIC)
 - Fixed effect model: 18200.0 for 40.2 parameters
 - Random effect model: 18199.4 for 44.3 parameters
 - Drop in DIC of -0.6 for 4.1 more parameters

Random effects

- Random effects capture the differences between the cities that weren't captured by covariates
- Random effects akin to residuals
- Also akin to data reduction (although in this example sample size of only 8)
- Plot random effects by covariates (diagnostic plots)

Results: seasonal amplitude by climate

- January (summer) temperature (left), July (winter) temperature (right)



Closing remarks

- Still important to checking the residuals
- Checking for remaining seasonality in this model showed a heat effect in elderly women in Brisbane and Sydney
- Future work needs more individual level data, compared to this ecological study
- The random effects model has identified some interesting hypothesis, but these need to be tested at an individual level

BCA/AEA Multilevel Modelling (MLM) workshop - Hobart, Sunday 26th August 2007

Venue: Hotel Grand Chancellor, 1 Davey Street, Hobart – Harbour View Room One, Level 2

PROGRAM

- 8.00-8.45 **Registration**
8.45-9.00 **Welcome and Introductory Remarks**

Introduction

- 9.00-9.45 **Introduction to MLMs** *Prof Alastair Leyland*, Public Health Sciences Unit,
University of Glasgow
9.45-10.30 **Fitting MLMs** *Prof Alastair Leyland*
10.30-11.00 Morning Tea

Applications (30 minute presentations with 5 minute discussion each and 15 minutes together at conclusion)

- 11.00-11.35 **Application #1** *Prof Alastair Leyland*
11.35-12.10 **Application #2** *Dr Adrian Barnett*, Population Health, University of Queensland
12.10-12.45 **Application #3** *A/Prof Anne Kavanagh*, Key Centre for Women's Health in Society,
University of Melbourne
12.45-1.00 **Discussion**
1.00-2.00 Lunch

All the world's a multilevel model

- 2.00-2.15 **Family and twin studies** *Dr Katrina Scurrah*, Dept Physiology, University of Melbourne
2.15-2.30 **Cluster randomised trials** *Dr Obi Ukoumunne*, Clinical Epidemiology and Biostatistics Unit,
Murdoch Childrens Research Institute, Royal Children's Hospital, Melbourne
2.30-2.45 **Population pharmacokinetic studies and nonlinear multilevel models**
Dr Julie Simpson, Centre for MEGA Epidemiology, University of Melbourne
2.45-3.00 **Bivariate mixed models for assessing change: An example**
A/Prof Andrew Forbes, Dept of Epidemiology & Preventive Medicine, Monash University
3.00-3.15 **Discussion**
3.15-3.45 Afternoon Tea

Special topics

- 3.45-4.00 **Software for multilevel modelling** *Dr Lyle Gurrin (presented by Dr Julie Simpson)*, Centre
for MEGA Epidemiology, University of Melbourne
4.00-4.30 **Recent Developments in MLMs** *Prof Alastair Leyland*
4.30-4.55 **Discussion**
4.55-5.00 **Closing remarks**